METRON

RIVISTA INTERNAZIONALE DI STATISTICA — REVUE INTERNATIONALE DE STATISTIQUE
INTERNATIONAL REVIEW OF STATISTICS — INTERNATIONALE STATISTISCHE ZEITSCHRIFT
REVISTA INTERNACIONAL DE ESTADISTICA

DIRETTORE PROPRIETARIO — DIRECTEUR ET PROPRIÉTAIRE
EDITOR AND PROPRIETOR — HERAUSGEBER UND EIGENTHÜMER
DIRECTOR Y PROPIETARIO

Prof. Dott. Corrado Gini, *della Università di Roma*

AMMINISTRATORE — ADMINISTRATEUR — MANAGER — VERWALTER — ADMINISTRADOR

Dott. Franco Mariani, *Istituto di Statistica della Università di Roma*

SEGRETARI DI REDAZIONE — SECRÉTAIRES DE RÉDACTION
EDITORIAL SECRETARIES — REDACTIONSSECRETAERE — SECRETARIOS DE LA REDACCIÓN

Dott. C. Benedetti, Prof. V. Castellano
Proff. N. Federici, E. Pizzetti, G. Pompilj

Vol. XVIII - N. 1-2 **30-IV-1956**

SOMMARIO-SOMMAIRE-CONTENTS-INHALT-SUMARIO

ROMA

AMMINISTRAZIONE DEL « METRON »

UNIVERSITÀ — ISTITUTO DI STATISTICA

ARTICOLI GIUNTI ALLA RIVISTA CHE VERRANNO PUBBLICATI NEI PROSSIMI NUMERI (*secondo l'ordine di arrivo*).

ARTICLES RECEIVED BY THE REVIEW WHICH WILL BE PUBLISHED IN FUTURE ISSUES (*according to date of receipt*).

ARTICULOS LLEGADOS A LA REVISTA QUE SE PUBLICARÁN EN LOS PROXIMOS NUMEROS (*según el orden de su llegada*).

ARTICLES REÇUS PAR LA REVUE ET À PARAÎTRE PROCHAINEMENT (*d'après la date de reception*).

ARTIKEL DIE AN DIE ZEITSCHRIFT ANGELANGT SIND UND WELCHE IN DEN FOLGENDEN NUMMERN ERSCHEINEN WERDEN (*nach der Reihenfolge des Eingangs*).

C. GINI. — *Sui limiti della invertibilità delle relazioni statistiche.*

O. M. J. MITTMANN. — *On the variance of the integral of an empirical function.*

J. C. KOOP. — *A general method of working out linear estimates of population totals from sample data of human populations on a tabulator.*

S. GATTI. — *Di una disuguaglianza tra momenti.*

S. GATTI. — *Sui limiti di alcune medie nel caso di successioni a più limiti.*

C. BENEDETTI. — *Di alcune disuguaglianze collegate ad indici statistici.*

CORRADO GINI

Généralisations et applications de la théorie de la dispersion [1]

1. La théorie de la dispersion fournit des critériums pour déterminer si les différences que les intensités d'un phénomène donné présentent dans divers espaces de temps, ou dans diverses circonscriptions territoriales, ou dans des catégories qui se distinguent par les modalités d'autres phénomènes, ont caractère accidentel ou systématique. Dans ce but, on détermine ordinairement un indice de variabilité des intensités qui se présentent dans les diverses catégories et on le compare à sa valeur théorique telle qu'elle serait par le seul effet du hasard.

Les premières applications de cette théorie — faites par Dormoy (1874), qui employait comme indice de variabilité l'écart simple moyen, et par Lexis (1876), qui employait l'écart quadra-

[1] Cet article présente les résultats de recherches poursuivies — avec quelques interruptions — pendant plus de 40 ans. J'ai eu l'idée de généraliser la théorie de la dispersion aux grandeurs absolues dès avant la première guerre mondiale, et les premières applications (v. p. 12-13) en ont été faites en 1915. Interrompues, d'abord à cause de ma participation à la guerre et ensuite à cause d'autres problèmes d'utilité immédiate sur lesquels s'était concentrée mon attention dans l'après-guerre, les recherches ont été reprises et pratiquement achevées avant la deuxième guerre mondiale.

Les principaux résultats au point de vue théorique ont été alors exposés dans la communication *Di una estensione della teoria della dispersione a serie di grandezze assolute*, présentée au Séminaire Statistique de l'Université de Rome dans la séance du 24 mai 1941 et résumée dans les Actes de la IV^e^ et de la V^e^ Réunion Scientifique de la Societé Italienne de Statistique (Rome, juin-juillet 1941), p. 405-406. Ils étaient destinés à être publiés dans l'« Archiv für Mathematische Wirtschaft- und Sozialforschung », mais cette Revue cessa de paraître pendant les dernières années de la guerre. Les recherches, interrompues de nouveau dans la période de l'après-guerre, ont été reprises en 1952 et les passages principaux du texte qui avait été préparé pour ladite Revue ont été publiés dans l'article *Estensione della teoria della dispersione e della*

tique moyen — se rapportaient à des fréquences relatives (par exemple : coefficients de mortalité, rapports des naissances de garçons au total des naissances, etc.).

Von Bortkiewicz étendit ces applications aux moyennes de grandeurs mesurables, telles que la stature, le revenu, etc.. On peut examiner, par exemple, si c'est accidentellement ou systématiquement que la stature moyenne des conscrits ou le revenu moyen des contribuables varie d'une année à l'autre dans le même pays, ou d'une circonscription à une autre, dans une certaine année.

2. Le schéma adopté pour déterminer la valeur théorique de l'indice de variabilité choisi était le même dans les deux cas.

En indiquant par $1, 2, 3, \ldots s$ les sections (espaces de temps, circonscriptions territoriales, catégories qualitatives) dans lesquelles on divise le champ d'observation, par $n_1, n_2, \ldots n_s$ et en général par n_k où $k = 1, 2, \ldots s$) le nombre des observations faites dans chacune d'elles, par $N = \sum_{k=1}^{s} n_k$ leur total : par $1, 2, \ldots, t$ les modalités du caractère (sexe, survivance, stature, etc.), par $M_1, M_2 \ldots M_t$ le nombre des fois où elles se présentent et par $N = M_1 + M_2 \ldots + M_t$ leur total correspondant au nombre total des observations, on suppose que dans une urne aient été mises N boules de t couleurs, correspondant en nombre aux t modalités possibles et l'on procède ensuite à extraction au hasard et sans remise, d'abord de n_1 boules que l'on attribue à la section 1, puis de n_2 boules que l'on attribue à la section 2 et ainsi de suite jusqu'à n_s boules que l'on attribue à la section s.

---

connessione a serie di grandezze assolute paru dans le « Giornale dell'Istituto Italiano degli Attuari », année XV[e] n. 1-2, 1952, p. 4-24).

D'autres résultats ont été publiés dans l'article *Estensioni e portata della teoria della dispersione*, inséré dans le volume *Studies in Mathematics and Mechanics* presented to *Richard von Mises* by Friends, Colleagues and Pupils, Academic Press, New York, 1954. Une erreur typographique assez sérieuse ayant echappé lors de l'impression, l'article a été reproduit dans le « Giornale dell'Istituto Italiano degli Attuari », année XVIII[e], 1955, p. 1-14. Nos recherches ont fait aussi l'objet d'une leçon en langue française donnée à l'Institut Henri Poincaré le 5 mai 1954, et qui vient d'être publiée dans les « Annales » dudit Institut. Tome XIV, fasc. IV, 1955. L'exposé fait dans cette leçon a été complété au point de vue théorique dans le présent article et accompagné de plusieurs applications des formules proposées qui montrent leur utilité pratique.

Ce *schéma* peut être appelé *des tirages au sort des modalités appartenant à chaque section* ou, plus brièvement, *schéma des tirages au sort.*

3. On peut avoir recours aussi à un autre schéma. Supposons que dans une urne on ait encore mis N boules, mais de s couleurs différentes correspondant aux sections, et en nombre égal à celui des observations faites dans les sections respectives, et que l'on tire de l'urne, au sort et sans répétition, d'abord M_1 boules que l'on attribue à la modalité 1, puis M_2 boules que l'on attribue à la modalité 2, et ainsi de suite jusqu'à M_t boules que l'on attribue à la modalité t.

Ce schéma peut être appelé *schéma des tirages au sort des sections entre lesquelles doivent être réparties au hasard les modalités* ou, plus brièvement, *schéma des répartitions au hasard.*

Les résultats que l'on obtient par les deux schémas sont identiques.

4. Si l'on indique par $a_{1k}, a_{2k} \ldots a_{n_k k}$ et en général par a_{ik} (où $i = 1, 2, \ldots n_k$) les valeurs observées dans les n_k observations relatives à la $k$ième section, par

$$A_k = \frac{1}{n_k} \sum_{i=1}^{n_k} a_{ik}$$

les moyennes des n_k valeurs observées dans la $k$ième section, et donc par

$$A = \frac{1}{N} \sum_{k=1}^{s} n_k A_k = \frac{\sum_{k=1}^{s} \sum_{i=1}^{n_k} a_{ik}}{N}$$

la moyenne des N valeurs observées dans toutes les k sections, l'écart quadratique moyen effectif des s valeurs de A_k sera

$$\sqrt{\frac{1}{N} \sum_{k=1}^{s} n_k (A_k - A)^2}$$

et l'écart quadratique moyen qu'il y aurait lieu de prévoir par l'effet du hasard serait

$$\sqrt{\frac{(s-1)\sum_{k=1}^{s}\sum_{i=1}^{t}(a_{ik}-A)^2}{(N-1)\,N}}$$

dont on tire l'indice de dispersion

$$Q_1=\sqrt{\frac{(N-1)\sum_{k=1}^{s}n_k\,(A_k-A)^2}{(s-1)\sum_{k=1}^{s}\sum_{i=1}^{t}(a_{ik}-A)^2}} \qquad (1)$$

On dit que Q_1 est un indice de *dispersion subnormale, normale* ou *supernormale* (1) selon qu'on a $Q_1 \lessgtr 1$.

Dans le cas où il s'agisse de grandeurs intensives, soit de caractères énumérables pour lesquels il n'y a que seulement deux intensités 1 ou 0, les moyennes A_k et A sont constituées par des rapports de fréquence que nous désignerons par m_k/n_k, M/N et la formule (1) peut être transformée ainsi

$$Q_2=\sqrt{\frac{(N-1)\sum_{k=1}^{s}n_k\left(\frac{m_k}{n_k}-\frac{M}{N}\right)^2}{(s-1)\,M\,(1-M/N)}} \qquad (2)$$

Il y a une différence importante entre les formules (1) et (2), au point de vue de leur applicabilité. Pour appliquer la formule (2) il suffit, en effet, de connaître (outre les valeurs de m_k et de n_k) la fréquence moyenne M/N ; pour appliquer la formule (1) il faut, par contre, connaître (outre les valeurs de A_k et de n_k) non seulement l'intensité moyenne A, mais aussi les écarts

(1) Peut-être serait-il préferable de parler, plutôt que de *dispersion subnormale, normale et supernormale*, de *dispersion hypobinomiale, binomiale et hyperbinomiale*, tout au moins lorsqu'il s'agit de grandeurs intensives, ainsi que je l'ai fait dans l'article *Asimmetria e anormalità delle distribuzioni statistiche* (dans cette même Revue, N. 1-2, 1951, p. 63 et suiv.). En effet la valeur M/N peut-être très différente de ½, de sorte que la distribution théorique peut être asymmétrique. De la sorte on écarterait le risque que, dans l'expression « dispersion normale », le mot « normale » puisse être entendu

$a_{ik} - A$ que présentent les intensités individuelles à partir de cette moyenne. Cela provient du fait que, pour les fréquences relatives, une fois que l'on connaît l'intensité moyenne M/N, on en déduit immédiatement l'écart quadratique moyen, tandis, que cela n'est pas vrai pour les autres moyennes.

5. Si les deux schémas considérés ci-dessus des tirages au sort et des répartitions au hasard conduisent au même résultat, le deuxième offre cependant un intérêt particulier, parce que, avec de légers ajustements, il peut être étendu au cas des phénomènes susceptibles de se répéter, auxquels le premier n'est, par contre, pas applicable.

Les deux schémas peuvent être appliqués, par exemple, pour juger si les décès se distribuent au hasard, dans un espace de temps donné, entre les habitants des diverses circonscriptions territoriales, ou, dans une circonscription territoriale donnée, entre les espaces de temps successifs, mais ils ne peuvent, par contre, être utilisés pour déterminer si, entre ces habitants, les rhumes se distribuent au hasard. Cela parce que, dans le premier cas, la somme des cas présentant les diverses modalités du phénomène est égale au total des individus, comme il est supposé dans les deux schémas ci-dessus considérés (par exemple la somme des morts et des survivants est égale au nombre des habitants exposés à mourir, étant donné que tout individu ou meurt ou survit et naturellement ne peut mourir qu'une seule fois) ; tandis que, dans le second cas, un individu peut avoir été ou non enrhumé dans un espace de temps, mais peut aussi avoir été plusieurs fois enrhumé dans cet espace de temps, de sorte que les rhumes plus les individus qui n'ont jamais été enrhumés ne sont pas équivalents au nombre des habitants exposés aux rhumes, et par conséquent les formules ci-dessus ne sont pas applicables.

---

dans le sens qu'on lui donne lorsqu'on parle de « courbe normale », c'est à dire de « courbe gaussienne ». Dans l'expression « courbe normale », le mot « normale « concerne la *forme* de la dispersion ; dans l'expression « dispersion normale », il concerne l'*intensité* de la dispersion. L'opportunité de distinguer les deux aspects avec des mots différents a été signalée dès 1908, dans notre ouvrage *Il sesso dal punto di vista statistico*, Palermo, Sandron, 1908 (maintenant en vente à la Revue « Metron », Rome, Via Terme di Diocleziano), p. 84-85.

Le schéma des répartitions au hasard devient cependant applicable si, au lieu de considérer des tirages au sort sans remise, nous considérons des tirages au sort avec remise.

Supposons avoir pour cela une urne contenant des boules en nombre égal à celui N des habitants du pays considéré, et de couleurs différentes selon les circonscriptions territoriales en lesquelles nous supposons que la population soit divisée. Nous nous demandons si les M rhumes qui se sont manifestés dans une année sont répartis accidentellement entre les habitants des circonscriptions. S'il s'agissait de décès au lieu de rhumes, nous pourrions, sur la base du schéma des répartitions au hasard ci-dessus considéré, tirer au sort de l'urne sans remise M boules qui seraient de couleurs diverses correspondant aux circonscriptions où les décès auraient lieu par l'effet du hasard. Mais nous ne pouvons pas en faire autant pour les M rhumes qui pourront être, et seront vraisemblablement, plus nombreux que les habitants et donc que les boules contenues dans l'urne. Du moment que le rhume peut se reproduire sur un même individu, nous devons considérer le cas où la boule qui a déjà été extraite, et correspond au rhume d'un individu donné, peut être extraite de nouveau en correspondance à un rhume ultérieur de cet individu.

Si nous indiquons par $m_1 , m_2 \ldots m_s$ les rhumes de s circonscriptions ayant respectivement $n_1 , n_2 , \ldots n_s$ habitants, la distribution des rhumes sera accidentelle lorsque les rapports $m_1/M , m_2/M \ldots m_s/M$ des rhumes dans les diverses circonscriptions aux rhumes de tout le pays considéré s'écarteront par des différences accidentelles des rapports $n_1/N , n_2/N \ldots n_s/N$ des habitants des diverses circonscriptions aux habitants de tout le pays.

Au lieu de comparer, comme dans les schémas précédents, des rapports théoriques à des rapports effectifs de dérivation, nous sommes amenés ainsi à comparer des rapports théoriques à des rapports effectifs de composition.

Nous appellerons ce schéma : *schéma de répartition au hasard sur la base des rapports de composition* ou simplement *schéma des rapports de composition* par opposition aux deux précédents, que nous appellerons *schémas des tirages au sort*, et, respectivement, *des répartitions au hasard sur la base des rapports de dérivation* ou simplement *schémas des rapports de dérivation*.

6. Selon le schéma des rapports de composition, le carré de l'écart effectif de la première circonscription sera $(m_1/M - n_1/N)^2$ et sa valeur théorique $(n_1/N)\ (1 - n_1/N)\ (1/M)$. Pour toutes les circonscriptions l'indice de dispersion sera donc :

$$Q_3 = \sqrt{\frac{M \sum_{i=1}^{s} \left(\frac{m_k}{M} - \frac{n_k}{N}\right)^2}{\sum_{k=1}^{s} \frac{n_k}{N}\left(1 - \frac{n_k}{N}\right)}} = \sqrt{\frac{M \sum_{k=1}^{s} \left(\frac{m_k}{M} - \frac{n_k}{N}\right)^2}{1 - \frac{1}{N^2} \sum_{k=1}^{s} n_k^2}} \qquad (3)$$

Dans le cas où n_k est constant $= n = N/s$, en faisant $M/s = m$, l'expression (3) devient :

$$Q'_3 = \sqrt{\frac{\sum_{k=1}^{s} (m_k - m)^2}{(s-1)\, m}} \qquad (3,1)$$

tandis que la formule (2) devient :

$$Q'_2 = \sqrt{\frac{N-1}{N} \frac{\sum_{k=1}^{s} (m_k - m)^2}{(s-1)\, m\, (1 - m/n)}} \qquad (2,1)$$

En faisant abstraction du facteur $(N-1)/N$, que l'on peut ordinairement négliger, il y a entre Q'_3 et Q'_2 le rapport très simple

$$Q'_3 = Q'_2 \sqrt{1 - \frac{m}{n}}$$

où n indique la limite maximum que m peut atteindre dans le cas de phénomènes reproductibles. Dans le cas où le phénomène est indéfiniment reproductible, $n = \infty$ et donc Q'_3 devient égal à Q'_2.

La formule (3,1) est également valable lorsqu'il s'agit de phénomènes très rares, de sorte que m peut être considéré négligeable vis-à-vis de n, quelle que soit la grandeur de celui-ci.

La diversité entre la formule (3,1) et la formule (2,1) est due au fait que, dans le schéma des rapports de composition, on admet que tous les M cas observés peuvent être attribués à un seul terme et, puisque le nombre M dépend du nombre des termes et rien n'empêche que l'on considère un nombre de termes de quel-

que grandeur que l'on veuille, cela revient à dire que l'on peut attribuer à un seul terme un nombre illimité de cas du phénomène.

Par contre, dans le schéma des rapports de dérivation on admet qu'il y a, pour chaque terme, un nombre limité de cas réalisés, dont le maximum est indiqué par le nombre n des cas possibles, d'entre lesquels on suppose tirés au sort les cas réalisés.

Or, il faut remarquer que, dans la théorie de la variabilité et de la concentration, l'hypothèse a déjà été prise en considération qu'il y ait une limite supérieure à la valeur des termes singuliers de la série, et il a été trouvé qu'il faut alors introduire dans les indices relatifs un coefficient de correction qui est précisément celui qu'il faut introduire dans la valeur de Q'_3 pour obtenir la valeur de Q'_2 [1].

7. Le schéma des rapports de composition étend notablement le champ d'application de la théorie de la dispersion. Sur la base des formules (1) et (2), cette théorie pouvait s'appliquer aux fréquences relatives (rapports de natalité, de nuptialité, de mortalité, de masculinité, etc.) et aux autres grandeurs relatives ayant au dénominateur des quantités énumérables, telles que les moyennes arithmétiques (par exemple, stature moyenne, poids moyen, revenu moyen, etc.) : autrement dit, aux grandeurs relatives énumérables ou mesurables rapportées à des grandeurs énumérables.

Le schéma des rapports de composition permet d'étendre les applications à toutes les grandeurs énumérables absolues, telles que maladies, incendies, tremblements de terre ou secousses sismiques, éruptions volcaniques, pluies, chutes de neige, averses de grêle, nombre des poissons pêchés, des têtes de gibier tuées, des diamants découverts et ainsi de suite [2].

---

[1] Voir l'article *Sul massimo degli indici di variabilità assoluta e sulle sue applicazioni agli indici di variabilità relativa ed al rapporto di concentrazione*, dans « Metron » vol. VIII, N. 3, février 1930, reproduit dans *Memorie di Metodologia statistica*, Milan, Giuffré, 1939. Une nouvelle édition de cet ouvrage, mise au jour par MM. E. Pizzetti et T. Salvemini vient de paraître par le soins de la Maison « Eredi V. Veschi » de Rome (Viale dell'Università 7).

[2] Pour le schéma des rapports de composition et pour les applications qu'il rend possibles, cf. notre article *Estensione della teoria della dispersione e della connessione a serie di grandezze assolute*, « Giornale dell'Istituto Italiano degli Attuari », année XV, n. 1-2, 1952, Rome, §§ 1-4.

Parmi les phénomènes que nous venons de mentionner, il y en a plusieurs que l'on peut considérer comme indéfiniment reproductibles, dans le sens qu'il n'y a aucune limite supérieure au nombre des cas qui peuvent se produire, sauf celle qui provient en pratique de la limitation de l'intervalle de temps considéré : tels les tremblements de terre ou les secousses des dits tremblements, les maladies, les éruptions volcaniques, les pluies, les chutes de neige ou les averses de grêle. D'autres — tels que les poissons pêchés, les têtes de gibier tuées, les diamants découverts — ne sont pas reproductibles dans le sens qu'un poisson ne peut être pêché qu'une fois, un canard ne peut être tué qu'une fois, un diamant ne peut être découvert qu'une fois ; mais chaque poisson pêché, chaque canard tué, chaque diamant découvert représente un cas qui s'est produit parmi le nombre inconnu, mais fini, de poissons, de canards, de diamants, exposés à être pêchés, tués, découverts, un nombre que l'on suppose, et que l'on a raison de supposer, très grand vis-à-vis du nombre des poissons, des canards, des diamants réellement pêchés, tués, découverts.

Pour la première catégorie de phénomènes, la formule (3.1) est applicable exactement ; pour la deuxième catégorie, elle peut être appliquée comme succédané de la formule (2.1), qu'il n'est pas possible d'appliquer à cause de notre ignorance de la valeur de n.

8. Voyons quelques applications.

Première application : Nombre des oiseaux tués

Cette application concerne les oiseaux tués dans les « valli » de la Vénétie. Les chiffres m'ont été fournis par les propriétaires des « valli » ou par les chasseurs qui les avaient louées dans la période qui a précédé la première guerre mondiale, lorsque les recherches exposeés dans cet article ont été entreprises.

Voici (tableau 1) les résultats pour une « valle » de la province de Venise dans les 19 saisons de chasse de 1896-97 à 1914-15. L'application de la formule (3.1) conduit à une valeur de $Q'_3 = 12{,}55$.

Pour les neufs dernières années précédant la première guerre mondiale on a des données pour 3 autres « valli » (cfr. tableau II), qui portent à des valeurs de $Q'_3 = 8{,}20$; 12,33 ; 8,99.

Tableau I

Nombre des oiseaux tués dans une « valle » de la Vénétie pendant la période 1896/97 - 1914/15

SAISONS DE CHASSE	NOMBRES DES OISEAUX TUÉS	SAISONS DE CHASSE	NOMBRES DES OISEAUX TUÉS
1	2	1	2
1896-97	1.750	1906-07	1.235
1897-98	2.813	1907-08	2.102
1898-99	2.472	1908-09	2.293
1899-1900	1.956	1909-10	2.284
1900-01	2.163	1910-11	2.240
1901-02	1.477	1911-12	2.956
1902-03	3.212	1912-13	2.160
1903-04	3.015	1913-14	3.402
1904-05	3.406	1914-15	2.943
1905-06	2.683	TOTAL	46.562

Tableau II

Nombre des oiseaux tués dans trois « valli » de la Vénetie (A,B,C) dans la période 1906-1914

ANNÉES	A	B	C	Totaux (A+B+C)
	1	2	3	4
1906	541	1.293	937	2.771
1907	661	955	836	2.452
1908	1.104	763	360	2.227
1909	832	942	515	2.289
1910	1.093	1.247	624	2.964
1911	1.144	1.532	463	3.139
1912	1.058	1.977	1.034	4.069
1913	1.332	1.590	859	3.781
1914	845	2.065	937	3.847
TOTAUX	8.610	12.364	6.565	27.539

Tableau III

Nombre des oiseaux tués dans la « valle » Marzotto (Vénétie) dans la période 1942/43 - 1952/53

SAISONS DE CHASSE	FOULQUES	OISEAUX « DE PLUME »	TOTAUX
1	2	3	4
1942/1943	4.035	5.153	9.188
1943/1944	3.346	2.869	6.215
1944/1945 [1]	—	—	—
1945/1946 [1]	—	—	—
1946/1947	4.024	560	4.584
1947/1948	5.428	3.548	8.976
1948/1949	5.479	7.249	12.728
1949/1950	2.186	3.389	5.575
1950/1951	2.101	1.105	3.206
1951/1952	1.189	861	2.050
1952/1953	3.163	1.326	4.489
TOTAUX	30.951	26.060	57.011

[1] La chasse a été suspendue à cause de la guerre.

Ces données suggèrent que, lorsqu'il n'y a pas de perturbations exceptionelles, l'indice de dispersion dans les « valli » de la Vénétie est à peu près = 10.

Lorsque, au contraire, des causes exceptionnelles de perturbation interviennent, la valeur de Q_3 peut monter bien au-dessus de ce chiffre. Considérons, par exemple, la « valle » Marzotto, dans la province de Venise, pendant la période 1942-43 à 1952-53. La chasse y a été suspendue, à cause de la guerre, pendant les deux saisons 1944-45 et 1945-46, de façon que l'on peut disposer des données seulement pour neuf saisons de chasse. C'est là une première circonstance exceptionnelle ; la seconde, encore plus importante, est que, dans les dernières années, des travaux extraordinaires ont été accomplis en vue de l'assainissement des marais, travaux qui évidemment ont dérangé le passage des oiseaux. On en voit clairement les effets dans les chiffres du tableau III, qui m'ont été aimablement communiqués par les propriétaires de la « valle ». La valeur de Q_3' est de 42,42 pour le total des oiseaux tués ; elle est de 25,09 pour les foulques et de 41,55 pour les oiseaux que les chasseurs appellent « oiseaux de plume », c'est-à-dire canards, sarcelles et autres semblables. Les foulques,

oiseaux plus grossiers et moins craintifs, ont été troublées beaucoup moins que les oiseaux de plume par les travaux d'assainissement.

Deuxième application : Nombre des thons pêchés.

Des séries de données bien plus intéressantes, à cause de la longue période qu'elles couvrent et des comparaisons qu'elles permettent, concernent le nombre des thons pêchés dans les trois madragues de Porto Paglia, Porto Scuso e Isola Piana, pendant les années de 1829 à 1953. Ces madragues sont situées près de l'extrémité sud-occidentale de la Sardaigne et sont très voisines l'une de l'autre. La période d'observation couvre 125 années ; les données font défaut pourtant pendant neuf années pour Porto Paglia, pendant deux années pour Porto Scuso et également pendant deux années pour Isola Piana. Pour 115 années les données sont complètes pour les trois madragues et peuvent être additionnées (voir pour les chiffres et leurs sources, le tableau IV).

La valeur de l'indice Q'_3 est de 32,32 pour Porto Paglia, de 35,37 pour Porto Scuso et de 33,43 pour Isola Piana. Elle est plus élevée — c'est-à-dire, que les pêches y sont plus irrégulières — pour Porto Scuso ; mais, somme toute, il y a une analogie remarquable entre les trois indices.

Pour les trois madragues ensemble, la valeur de Q'_3 atteint 52,10. Il y a une différence très sensible entre cette valeur et celle, qui tourne autour de 33, pour chacune des madragues. Nous reviendrons sur ce point.

La longueur de la période considérée suggère de calculer les valeurs de Q'_3 pour des séries partielles. Nous avons calculé cette valeur d'abord pour des séries les plus courtes possibles, c'est-à-dire de deux années (par exemple 1929-30, 1931-32 etc.), après pour des séries de dix années et ensuite de 20, de 30 et de 40 années.

Les tableaux suivants V-VIII résument les résultats. Les distributions des valeurs de Q'_3 qu'ils mettent en évidence sont analogues pour les quatre tableaux. Elles présentent deux caractéristiques : d'un côté, la variabilité des valeurs de Q'_3 diminue, ainsi qu'il est naturel, avec l'augmentation du nombre des an-

Tableau IV

Nombres des thons pêchés dans les madragues de Porto-Paglia, Porto-Scuso et Isola Piana (Sardaigne) dans la période 1829-1953 (1)

ANNÉES	Porto Paglia	Porto Scuso	Isola Piana	Totaux pour les trois madragues
	1	2	3	4
1829	2.478	3.429	2.839	8.746
1830	3.030	3.285	1.477	7.792
1831	2.338	2.370	1.665	6.373
1832	2.074	3.701	640	6.415
1933	1.901	897	455	3.253
1834	2.385	1.660	970	5.015
1835	1.989	1.866	1.418	5.273
1836	1.331	1.900	1.114	4.345
1837	1.474	3.581	2.521	7.576
1838	2.753	3.983	2.753	9.489
1839	2.317	5.221	3.495	11.033
1840	3.171	6.307	4.669	14.147
1841	2.373	3.585	2.935	8.893
1842	2.625	4.480	2.399	9.504
1843	2.362	3.314	2.527	8.203
1844	1.910	2.643	1.831	6.384
1845	1.878	3.058	1.464	6.400
1846	1.163	1.430	1.218	3.811
1847	433	760	141	1.334
1848	1.307	2.258	1.200	4.765
1849	4.085	5.458	3.188	12.731
1850	2.180	1.945	1.300	5.425
1851	3.474	3.627	2.407	9.508
1852	1.507	1.316	1.981	4.804
1853	2.658	3.223	3.323	9.204
1854	4.140	6.025	4.284	14.449
1855	2.550	3.885	2.381	8.816
1856	1.860	2.632	1.891	6.383
1857	2.780	3.608	3.150	9.538
1858	1.010	1.000	1.486	3.496
1859	3.550	3.680	3.207	10.437
1860	4.763	4.852	3.682	13.297
1861	1.907	4.457	2.160	8.524
1862	3.234	2.974	2.970	9.178
1863	4.511	4.970	2.772	12.253
1864	4.436	4.572	3.940	12.948
1865	7.168	5.007	4.564	16.739
1866	4.520	5.107	3.638	13.265
1867	2.427	4.236	3.231	9.894
1868	6.766	6.320	3.718	16.804
1869	4.745	5.633	4.811	15.189
1870	3.692	5.324	3.282	12.298
1871	3.017	4.020	3.596	10.633
1872	4.780	7.175	5.208	17.163

ANNÉES	PORTO PAGLIA	PORTO SCUSO	ISOLA PIANA	TOTAUX POUR LES TROIS MADRAGUES
	1	2	3	4
1873	3.041	7.275	5.150	15.466
1874	—	3.348	—	—
1875	1.910	2.461	2.745	7.116
1876	3.113	7.656	4.571	15.340
1877	3.515	6.441	3.857	13.813
1878	7.171	10.061	5.827	23.059
1879	2.153	—	3.002	—
1880	3.682	6.264	7.240	17.186
1881	5.618	8.000	5.850	19.468
1882	6.526	10.136	5.780	22.442
1883	3.990	6.928	5.800	16.718
1884	3.178	8.446	4.991	16.615
1885	1.573	3.656	2.075	7.304
1886	3.030	5.144	3.390	11.564
1887	2.226	4.776	4.372	11.374
1888	2.057	4.135	3.094	9.286
1889	2.102	4.707	3.704	10.513
1890	772	2.380	1.928	5.080
1891	1.890	5.144	1.870	8.904
1892	1.990	3.968	3.800	9.758
1893	316	1.587	1.603	3.506
1894	1.964	3.506	1.755	7.225
1895	413	1.609	1.636	3.658
1896	1.245	2.775	2.335	6.355
1897	2.774	6.055	3.544	12.373
1898	2.738	6.247	5.900	14.885
1899	1.714	3.240	2.781	7.735
1900	—	7.508	7.716	—
1901	—	9.478	5.281	—
1902	—	5.644	4.542	—
1903	1.609	4.600	6.227	12.436
1904	635	1.831	3.787	6.253
1905	673	4.478	5.508	10.659
1906	1.521	4.600	4.564	10.685
1907	850	3.292	2.090	6.232
1908	—	3.412	4.171	—
1909	—	4.776	8.404	—
1910	958	9.726	8.868	21.552
1911	928	7.569	9.792	18.289
1912	1.287	3.640	3.757	8.684
1913	724	2.866	4.142	7.732
1914	—	3.823	4.029	—
1915	1.875	3.999	3.788	9.662
1916	912	4.023	2.699	7.634
1917	729	2.548	1.812	5.089
1918	755	2.913	2.855	6.523
1919	772	2.986	2.694	6.452
1920	948	2.497	2.242	5.687

ANNÉES	PORTO PAGLIA	PORTO SCUSO	ISOLA PIANA	TOTAUX POUR LES TROIS MADRAGUES
	1	2	3	4
1921	322	867	1.256	2.445
1922	1.725	3.602	1.488	6.815
1923	1.932	3.897	2.865	8.694
1924	988	2.383	2.260	5.631
1925	1.808	2.072	1.544	5.424
1926	415	1.412	1.680	3.507
1927	794	1.533	759	3.086
1928	776	1.139	1.024	2.939
1929	933	1.214	1.146	3.293
1930	592	957	868	2.417
1931	688	1.172	899	2.759
1932	847	1.302	1.482	3.631
1933	192	586	879	1.657
1934	—	2.609	2.357	—
1935	539	1.171	786	2.496
1936	1.004	2.672	2.056	5.732
1937	1.448	2.402	2.127	5.977
1938	512	1.159	1.289	2.960
1939	734	1.482	1.208	3.424
1940	414	1.671	892	2.977
1941	1.917	3.363	2.470	7.750
1942	1.678	1.873	1.747	5.298
1943	—	—	—	—
1944	2.681	3.072	2.201	7.954
1945	731	1.288	1.376	3.395
1946	1.773	2.533	2.359	6.665
1947	1.109	2.559	2.630	6.298
1948	538	1.133	1.460	3.131
1949	1.467	1.889	1.718	5.074
1950	3.046	3.567	1.758	8.371
1951	1.105	2.794	2.564	6.463
1952	1.138	1.059	638	2.835
1953	304	676	450	1.430

(1) Les chiffres ont été puisés aux sources officielles. Pour les années 1829-1916, ils ont été publiés dans mon rapport *Ancora sulle statistiche delle tonnare di Sardegna*, Roma, Tipografia della Camera dei Deputati, 1917, p. 21-23. Pour les années suivantes, ils m'ont été aimablement communiqués par le Capitaine B. R BRUNO, chef de l'« Ufficio Circondariale Marittimo » de Carloforte (lettres du 25 mars et du 23 juillet 1953).

nées considérées ; de l'autre, le centre de la distribution se déplace, avec ladite augmentation du nombre des années considérées, vers les valeurs les plus élevées.

Cette deuxième circonstance peut être mesurée par l'augmentation de la valeur de la médiane, qui dans chaque distribution est marquée par une flèche, ou par celle de la moyenne

arithmétique, qui se trouve indiquée au bas de chaque distribution.

L'augmentation présente pourtant une exception très régulière : pour les séries de trente années, ladite moyenne, ainsi que généralement la médiane, est moindre que pour les séries correspondantes de vingt années : ce résultat fait penser que, dans les variations du nombre des thons pêchés, il y aurait un cycle qu'il serait intéressant d'étudier.

Pour les séries minima de deux années, la valeur de Q'_3 est, pour chaque madrague, contenue entre 15 et 19 tandis que, pour l'ensemble des trois madragues, elle dépasse 24. C'est la dispersion à attribuer aux variations annuelles de la pêche. L'augmentation de la valeur de Q'_3 avec l'extension de la série fait comprendre que, à côté des variations annuelles, il y a aussi des tendances systématiques dont l'effet est d'autant plus prononcé que la période considérée est plus longue.

Il ressort de tout cela que l'on ne peut pas parler de la dispersion d'un phénomène sans préciser, non seulement les sections (intervalles de temps, circonscriptions, etc.) dans lesquelles les données sont classées, mais aussi le nombre de ces sections, c'est-à-dire la longueur de la série.

Les résultats exposés dans les tableaux V-VIII prêtent à d'autres considérations intéressantes.

Les moyennes des valeurs de Q'_3 sont toujours plus élevées pour les séries des pêches globales concernant les trois madragues que pour les séries respectives concernant chacune des trois madragues. Cette différence provient de la solidarité qui existe entre les variations que les pêches présentent d'une année à l'autre dans les trois madragues.

La solidarité s'explique par le fait que les trois madragues, qui ne sont pas très éloignées l'une de l'autre, exploitent, plus ou moins intensivement, les mêmes troupes de thons qui se dirigent vers le détroit qui sépare l'ile de Saint-Pierre de la côte occidentale de la Sardaigne.

Il y a lieu aussi de remarquer que la différence signalée ci-dessus est moins forte lorsque l'on considère les séries partielles de deux années que lorsque l'on considère les séries partielles plus longues ou bien les séries entières de 115 termes.

Tableau V

Porto Paglia : Thons pêchés

Valeurs de l'indice de dispersion Q'_3	Nombres des indices de dispersion avant la valeur indiquée à la col. 1 pour les séries de				
	2 années	10 années	20 années	30 années	40 années
1	2	3	4	5	6
0-2	6	—	—	—	—
2-4	3	—	—	—	—
4-6	2	—	—	—	—
6-8	4	—	—	—	—
8-10	4	—	—	—	—
10-12	5	1	—	—	—
12-14	2	1(1)	—	—	—
14-16	1	1(2)	1	—	—
16-18	→ 3	1	—	1	—
18-20	5	1	1(3)	1(4)	1(5)
20-22	7	→2(6)	→2(7)	→1(8)	—
22-24	1	2(9)	—	—	—
24-26	4	2(10)	1(11)	1(12)	—
26-28	1	1	1	—	→1
28-30	2	—	—	—	—
30-32	1	—	—	—	—
32-34	—	—	—	—	1(13)
34-36	2	—	—	—	—
36-38	—	—	—	—	—
38-40	—	—	—	—	—
40-42	—	—	—	—	—
42-44	—	—	—	—	—
44-46	1	—	—	—	—
46-48	—	—	—	—	—
48-50	—	—	—	—	—
50-52	—	—	—	—	—
Totaux	54	12	6	4	3
Moyennes arithmétiques des valeurs de Q'_3	15,45	19,79	21,37	20,77	26,31

(1) Série de 9 années.
(2) Série de 6 années.
(3) Série de 18 années.
(4) Série de 28 années.
(5) Série de 36 années.
(6) 1 série de 9 années.
(7) 1 série de 16 années et 1 série de 18 années.
(8) Série de 24 années.
(9) 1 série de 8 années.
(10) 1 série de 9 années.
(11) Série de 19 années.
(12) Série de 29 années.
(13) Série de 35 années.

Tableau VI

Porto Scuso : Thons pêchés

Valeurs de l'indice de dispersion Q'_3	Nombres des indices de dispersion ayant la valeur indiquée à la col. 1 pour les séries de				
	2 années	10 années	20 années	30 années	40 années
1	2	3	4	5	6
0-2	7	—	—	—	—
2-4	3	—	—	—	—
4-6	4	—	—	—	—
6-8	3	—	—	—	—
8-10	2	—	—	—	—
10-12	3	—	—	—	—
12-14	3	1	—	—	—
14-16	4	—	—	—	—
16-18	→3	1(1)	—	—	—
18-20	1	1	1(2)	—	—
20-22	4	2	—	1(3)	—
22-24	2	—	1	—	—
24-26	2	—	1	—	1
26-28	4	→2	→ 1	→2(3)	—
28-30	3	3	—	—	—
30-32	1	—	1	1	→1(4)
32-34	1	2	—	—	1(4)
34-36	2	—	—	—	—
36-38	1	—	1	—	—
38-40	2	—	—	—	—
40-42	4	—	—	—	—
42-44	—	—	—	—	—
44-46	—	—	—	—	—
46-48	—	—	—	—	—
48-50	—	—	—	—	—
50-52	1	—	—	—	—
Totaux	60	12	6	4	3
Moyennes arithmétiques des valeurs de Q'_3	18,61	25,07	27,12	26,07	30,57

(1) Série de 9 années.
(2) Série de 19 années.
(3) Série de 29 années.
(4) Série de 39 années.

Tableau VII

Isola Piana: *Thons pêchés*

Valeurs de l'indice de dispersion Q'_3	Nombres des indices de dispersion, ayant la valeur indiquée à la col. 1 pour les séries de				
	2 années	10 années	20 années	30 années	40 années
1	2	3	4	5	6
0-2	3	—	—	—	—
2-4	4	—	—	—	—
4-6	4	—	—	—	—
6-8	10	—	—	—	—
8-10	2	—	—	—	—
10-12	6	1	—	—	—
12-14	→4	—	—	—	—
14-16	4	3(1)	1(2)	1(3)	—
16-18	4	1	1	—	—
18-20	1	1	—	1(3)	—
20-22	4	→ 1	1	→	—
22-24	2	1	→	1	1
24-26	5	1	1	—	→1(4)
26-28	1	2	—	—	—
28-30	2	—	1	—	—
30-32	—	—	—	—	—
32-34	—	—	—	—	—
34-36	1	—	—	1	—
36-38	—	—	—	—	—
38-40	—	—	—	—	—
40-42	1	1	—	—	—
42-44	—	—	—	—	1(4)
44-46	—	—	1	—	—
46-48	—	—	—	—	—
48-50	1	—	—	—	—
50-52	1	—	—	—	—
Totaux	60	12	6	4	3
Moyennes arithmétiques des valeurs de Q'_3	15,00	21,46	25,37	23,50	30,11

(1) Deux séries de 9 années.
(2) Série de 19 années.
(3) Série de 29 années.
(4) Série de 39 années.

Tableau VIII

Porto Scuso, Porto Paglia et Isola Piana : Thons pêchés

VALEURS DE L'INDICE DE DISPERSION Q'_3	NOMBRES DES INDICES DE DISPERSION AYANT LA VALEUR INDIQUÉE À LA COL. 1 POUR LES SÉRIES DE				
	2 années	10 années	20 années	30 années	40 années
1	2	3	4	5	6
0-2	4	—	—	—	—
2-4	—	—	—	—	—
4-6	4	—	—	—	—
6-8	3	—	—	—	—
8-10	1	—	—	—	—
10-12	2	—	—	—	—
12-14	1	—	—	—	—
14-16	6	—	—	—	—
16-18	1	—	—	—	—
18-20	→ 4	—	—	—	—
20-22	3	—	—	—	—
22-24	—	—	—	—	—
24-26	2	2(1)	—	—	—
26-28	1	2(2)	—	—	—
28-30	1	2(1)	1(3)	1(4)	—
30-32	2	—	—	—	—
32-34	2	→ —	—	—	—
34-36	3	—	2(5)	1(4)	—
36-38	1	2	→ 1	→ 1	—
38-40	2	—	1(3)	—	—
40-42	—	2	—	—	1
42-44	3	1	—	—	—
44-46	2	—	—	1(6)	—
46-48	—	—	—	—	→ 1(7)
48-50	1	—	—	—	—
50-52	—	—	—	—	—
52-54	1	—	—	—	1(8)
54-56	2	—	—	—	—
56-58	—	1(9)	1(3)	—	—
58-60	1	—	—	—	—
TOTAL	53	12	6	4	3
Moyennes arithmétiques des valeurs de Q'_3	24,15	34,72	38,78	36,61	46,47

(1) 1 Série de 9 années.
(2) 1 série de 6 années.
(3) Série de 18 années.
(4) Série de 28 années.
(5) 1 série de 16 années.
(6) Séries de 24 années.
(7) Séries de 34 années.
(8) Séries de 36 année.
(9) Séries de 8 année.

En effet, pour les séries de deux années, les valeurs de Q_3' sont = 15,45 pour Porto Paglia, = 15,00 pour Isola Piana et = 18,61 pour Porto Scuso, avec une moyenne de 16,35, contre une valeur de Q_3' = 24,15 pour les trois madragues prises ensemble, ce qui représente une augmentation de 47,7 %. Pour les séries successives, l'augmentation respective est de 57,0 % pour les séries de dix années, de 57,5 % pour les séries de vingt années, de 56,1 % pour les séries de trente années, de 60,2 % pour les séries de quarante années, de 54,7 % pour les séries entières de 115 années.

Il y a lieu de conclure, d'après ces chiffres, que la solidarité que nous avons signalée résulte de deux facteurs : une solidarité dans les variations annuelles, qui se manifeste dans les séries de deux années, et une solidarité dans les tendances systématiques des pêches, qui s'ajoute à la première dans les séries plus longues.

Ces résultats sont confirmés par les rapports de corrélation (formule de Bravais) que nous avons calculés entre les thons pêchés dans des trois madragues prises deux à deux, en considérant d'abord les séries de deux années et ensuite les séries de quarante années.

CORRÉLATION ENTRE LES MADRAGUES DE	SÉRIES DE	
	2 années	40 années
Porto Scuso et Porto Paglia.	0,68	0,74
Porto Scuso et Isola Piana	0,66	0,79
Porto Paglia et Isola Piana	0,45	0,54

Ces chiffres montrent que la corrélation est plus élevée entre les madragues de Porto Scuso et de Porto Paglia et entre celles de Porto Scuso et d'Isola Piana, qui sont plus proches l'une de l'autre, qu'entre celles de Porto Paglia et d'Isola Piana, séparées par une distance un peu plus grande.

On remarque aussi que la corrélation est toujours moins élevée pour les séries de deux années que pour celles de qua-

Nombres des tremblements de te

ANNÉES	NOMBRES *a*) des tremblements de terre *b*) des secousses sismiques	NOMBRES DES SECOUS														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1917	*a*) 164 *b*) *560*	54 *54*	38 *76*	19 *57*	13 *52*	14 *70*	10 *60*	5 *35*	2 *16*	2 *18*			1 *12*	1 *13*		1 *15*
1918	*a*) 97 *b*) *266*	45 *45*	18 *36*	12 *36*	2 *8*	9 *45*	3 *18*	5 *35*			1 *10*		1 *12*			
1919	*a*) 81 *b*) *193*	40 *40*	16 *32*	4 *12*	11 *44*	3 *15*	4 *24*		2 *16*		1 *10*					
1920	*a*) 69 *b*) *172*	32 *32*	14 *28*	10 *30*	6 *24*		3 *18*	2 *14*		1 *9*						
1921	*a*) 70 *b*) *214*	35 *35*	11 *22*	3 *9*	4 *16*	1 *5*	6 *36*	5 *35*		2 *18*	1 *10*		1 *12*			
1922	*a*) 69 *b*) *136*	43 *43*	9 *18*	8 *24*	4 *16*		2 *12*	2 *14*		1 *9*						
1923	*a*) 88 *b*) *163*	55 *55*	15 *30*	6 *18*	7 *28*	1 *5*	2 *12*	1 *7*	1 *8*							
1924	*a*) 74 *b*) *214*	28 *28*	19 *38*	9 *27*	9 *36*	5 *25*	1 *6*	1 *7*				1 *11*				
1925	*a*) 89 *b*) *192*	53 *53*	13 *26*	9 *27*	6 *24*	1 *5*	2 *12*	2 *14*	1 *8*			1 *11*	1 *12*			
1926	*a*) 77 *b*) *135*	54 *54*	12 *24*	5 *15*	2 *8*	1 *5*		1 *7*	1 *8*						1 *14*	
1927	*a*) 99 *b*) *334*	35 *35*	25 *50*	13 *39*	8 *32*	3 *15*	4 *24*	5 *35*		2 *18*						
1928	*a*) 161 *b*) *676*	57 *57*	25 *50*	25 *75*	16 *64*	7 *35*	4 *24*	7 *49*		4 *36*	2 *20*	2 *22*	1 *12*	2 *26*	1 *14*	1 *15*
1929	*a*) 117 *b*) *266*	55 *55*	29 *58*	10 *30*	14 *56*	7 *35*									1 *14*	
1930	*a*) 118 *b*) *296*	73 *73*	17 *34*	11 *33*	5 *20*	1 *5*	1 *6*		2 *16*	3 *27*	1 *10*		1 *12*			1 *15*
1917 1930	*a*) 1.373 *b*) *3.817*	659 *659*	261 *522*	144 *432*	107 *428*	53 *265*	42 *252*	36 *252*	9 *72*	15 *135*	6 *60*	4 *44*	6 *72*	3 *39*	3 *42*	3 *45*

Tableau IX

cousses sismiques en Italie (1917-1930)

TREMBLEMENT DE TERRE																		Nombre moyen des secousses pour chaque tremblement de terre
18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
1 *18*	1 *19*			1 *22*	1 *23*														3,41
			1 *21*																2,74
																			2,38
																			2,49
																			3,06
																			1,97
																			1,85
																		1 *36*	2,89
																			2,16
																			1,75
1 *18*																	1 *35*		3,37
		1 *20*	1 *21*		1 *23*	1 *24*		1 *26*					1 *31*	1 *32*					4,20
1 *18*																			2,27
1 *18*									1 *27*										2,51
4 *72*	1 *19*	1 *20*	2 *42*	1 *22*	2 *46*	1 *24*		1 *26*	1 *27*				1 *31*	1 *32*			1 *35*	1 *36*	2,78

rante années, conformément à ce que suggérait la comparaison entre les valeurs de Q'_3 pour les dites séries. C'est qu'à la solidarité entre les variations annuelles s'ajoute celle qui existe entre les tendances systématiques.

On peut remarquer aussi que la différence entre le coefficient de corrélation pour les séries de deux années et celui pour les séries de quarante années est plus prononcée entre les madragues de Porto Scuso et d'Isola Piana qu'entre les autres madragues, ce qui suggère une solidarité plus accentuée entre les tendances systématiques de ces deux madragues qu'entre celles des autres. Cette conclusion est conforme aux résultats d'une étude analytique que nous avons eu l'occasion de faire il y a presque une quarantaine d'années sur les pêches des madragues de Sardaigne de 1829 à 1916 (1).

Troisième application : Nombre des tremblements de terre.

En Italie, on a publié le nombre des tremblements de terre dans les années de 1891 à 1930 (2). L'application de la formule (3.1) à la serie complète, ainsi qu'aux deux séries partielles de vingt années et aux quatre séries partielles de dix années, dans lesquelles elle peut être divisée, donne les valeurs suivantes de l'indice de dispersion Q'_3 :

1891-00	2,95	2,62	3,81
1902-10	2,39		
1911-20	3,64	3,00	
1921-30	1,90		

Ici aussi, comme dans le cas des thons pêchés, on remarque que les séries longues tendent à présenter des valeurs de Q'_3 plus élevées que les séries courtes. La valeur 3,81 pour la série complète dépasse les valeurs pour les séries partielles, mais

---

(1) *Analisi delle statistiche sul prodotto delle tonnare di Sardegna*, Roma, Tipografia della Camera dei Deputati, 1916 ; *Ancora sulle statistiche delle tonnare di Sardegna*, Roma, ibidem, 1917.

(2) A. Cavasino, *La sismicità dell'Italia nel quarantennio 1891-1930*, « Bollettino della Società sismologica italiana », vol. XXX, 1931-1932, n. 5.

la différence paraît moins forte que dans l'application précédente. Cela peut être interprété dans le sens que les tendances systématiques dans les variations du nombre des tremblements de terre sont moins importantes que dans les variations du nombre des thons pêchés.

Il paraît très difficile de faire une statistique exacte des tremblements de terre. En effet les données portées par l'article cité ci-dessus étaient officielles. D'autres données pour les années 1917-1930, (voir tableau IX) m'ont été communiquées par le Bureau Central de Météorologie et d'Ecologie agraire, qui a la charge desdites statistiques. Bien qu'étant aussi officielles, ces données ne s'accordent pas avec les précédentes. La valeur de $Q'_3 = 3{,}20$, que l'on en tire, ne diffère pourtant pas sensiblement des résultats précédents.

Les séries des tremblements de terre paraissent donc caractérisées par un indice de dispersion qui s'approche de 3.

Quatrième application : Incendies.

Des statistiques détaillées sur les incendies pendant un siècle ont été publiées en Suisse en 1907 et en 1908 par les Instituts d'Assurance contre l'incendie du Canton d'Argovie et du Canton de Berne (1). Dans cette application, le nombre des cas d'incendie peut bien être rapporté aux nombre des bâtiments assurés, de façon que la formule à appliquer est la formule (2).

Pour le Canton de Berne, on obtient les valeurs suivantes de Q_2. Nous avons gardé la division en périodes faite par la publication officielle sur la base des variations de la législation. Ces variations empêchent de réunir les dites périodes en une période unique d'un siècle.

1807-34.	1,36
1835-58.	3,06
1859-82.	3,91
1883-906.	2,39
moyenne arithmétique	2,68

(1) Voir : *Die Aargauische Brandversicherungsanstalt, 1806-1906*, H. R. Sauerländer und Co., Aarau, 1907, et *Festschrift der Brandversicherungsanstalt des Kantons Bern*, Rösch und Schatzmann, Bern, 1908.

Pour le Canton d'Argovie, j'ai partagé la période en deux demi-siècles. Les valeurs de Q_2 sont les suivantes :

1806-55.	2,92
1856-905.	2,63

Elles s'approchent beaucoup de la moyenne (2,68) obtenue pour le Canton de Berne. Pour la série entière de l'Argovie, la valeur de Q_2 est sensiblement supérieure ($= 3{,}13$), ce qui témoigne d'une tendance systématique dans les variations.

Pour l'Argovie, la publication fournit aussi des données spéciales pour les cas d'incendies graves, qui sont naturellement beaucoup plus rares (336 dans le siècle, contre un total de 4.400 incendies). Ces cas ont été groupés par dizaines d'années et rapportés, pour chaque décade, au total des édifices assurés dans les dix ans. La valeur de Q_2 est $= 3{,}14$, en coïncidence presque parfaite avec celle obtenue pour tous les incendies.

Cinquième application : Jours de pluie.

Nous avons une très belle série de données pour les jours de pluie à Padoue ; elle couvre plus de deux siècles, allant de 1725 à 1940. Les données jusqu'à 1934 ont été publiées — après avoir été tirées de différentes sources et rendues comparables (autant qu'on le peut dans cette matière) — par le professeur G. Crestani, Directeur de l'Observatoire Météorologique G. Magrini de Padoue [1] ; elles ont été complétées jusqu'à 1940 par des données puisées au même Observatoire.

Nous les avons partagées en séries de 2, 5, 10, 20, 30, 40, 55, 80 années et avons également considéré la série de 210 années qui va de 1731 à 1940. Le tableau X contient la classification des valeurs de Q_2' que l'on a obtenues.

On remarque les mêmes caractéristiques que nous avons signalées pour les séries des pêches du thon. En passant de la colonne concernant les séries minima à deux termes à celles concernant les séries avec un nombre de termes plus élevé, la

---

[1] G. Crestani, F. Ramponi, L. Venturelli, *Le precipitazioni atmosferiche a Padova*, Magistrato delle Acque, Ufficio idrografico, pubbl. n. 137, 1935.

Tableau X

Ville de Padoue: Fréquence annuelle des jours de pluie (1725-1940)

VALEURS DE L'INDICE DE DISPERSION Q_2' 1	NOMBRES DES INDICES DE DISPERSION AYANT LA VALEUR INDIQUÉE À LA COL. 1 POUR LES SÉRIES DE								
	2 années 2	5 années 3	10 années 4	20 années 5	30 années 6	40 années 7	55 années 8	80 années 9	210 années 10
0, –0,1	5	—	—	—	—	—	—	—	—
0,1–0,2	6	—	—	—	—	—	—	—	—
0,2–0,3	13	—	—	—	—	—	—	—	—
0,3–0,4	7	—	—	—	—	—	—	—	—
0,4–0,5	5	1	—	—	—	—	—	—	—
0,5–0,6	8	—	—	—	—	—	—	—	—
0,6–0,7	6	2	—	—	—	—	—	—	—
0,7–0,8	→4	4	—	—	—	—	—	—	—
0,8–0,9	7	2	1	—	—	—	—	—	—
0,9–1,0	6	2	1	—	—	—	—	—	—
1,0–1,1	1	1	1	—	—	—	—	—	—
1,1–1,2	6	4	—	1	—	—	—	—	—
1,2–1,3	6	1	—	1	—	—	—	—	—
1,3–1,4	3	2	3	—	—	—	—	—	—
1,4–1,5	1	→3	3	1	1	1	—	—	—
1,5–1,6	4	3	→5	→4	1	→2	1	—	—
1,6–1,7	2	3	2	1	1	—	—	—	—
1,7–1,8	—	1	2	—	—	—	—	—	—
1,8–1,9	2	1	—	1	→2	—	—	—	—
1,9–2,0	—	2	1	—	1	—	1	—	—
2,0–2,1	—	4	1	—	—	1	→ —	1	—
2,1–2,2	—	—	1	—	—	—	1	→1	1
2,2–2,3	2	1	—	—	—	—	1	—	—
2,3–2,4	2	4	—	—	—	1	—	—	—
2,4–2,5	—	—	—	—	—	—	—	—	—
2,5–2,6	4	—	—	—	1	—	—	—	—
2,6–2,7	1	—	—	—	—	—	—	—	—
2,7–2,8	—	1	—	—	—	—	—		—
2,8–2,9	1	—	—	—	—	—	—	—	—
3,1–3,2	—	—	—	1	—	—	—	—	—
3,6–3,7	1	—	—	—	—	—	—	—	—
3,8–3,9	1	—	—	—	—	—	—	—	—
4,0–4,1	1	—	—	—	—	—	—	—	—
4,4–4,5	1	—	—	—	—	—	—	—	—
4,5–4,6	1	—	—	—	—	—	—	—	—
5,8–5,9	—	1	—	—	—	—	—	—	—
11,0–11,1	1	—	—	—	—	—	—	—	—
TOTAUX . . .	108	43	21	10	7	5	4	2	1
Moyennes arithmétiques des valeurs de Q_2'	1,14	1,56	1,51	1,67	1,83	1,80	1,97	2,09	2,17

Tableau XI

Ville de Padoue: Fréquence mensuelle des jours de pluie (1725-1930)

VALEURS DE L'INDICE DE DISPERSION Q_2' 1	Nombres des indices de dispersion ayant la valeur indiquée à la col. 1 pour le mois de											
	Janvier 2	Février 3	Mars 4	Avril 5	Mai 6	Juin 7	Juillet 8	Août 9	Septembre 10	Octobre 11	Novembre 12	Décembre 13
0,66–0,70	—	—	—	—	—	1	—	—	—	—	—	—
0,71–0,75	—	—	—	—	—	—	—	—	—	—	—	—
0,76–0,80	—	—	—	—	—	1	—	—	—	—	—	—
0,81–0,85	—	—	—	—	—	1	1	—	1	—	1	—
0,86–0,90	—	—	—	—	—	1	1	—	—	1	—	—
0,91–0,95	—	—	—	1	—	1	1	4	—	—	—	—
0,96–1,00	—	—	—	—	2	1	2	3	1	—	—	—
1,01–1,05	—	—	1	—	—	2	2	—	—	1	1	—
1,06–1,10	—	—	—	1	2	—	2	1	1	2	1	—
1,11–1,15	—	—	—	1	—	1	—	2	—	1	—	1
1,16–1,20	1	—	2	1	—	1	1	1	2	—	1	1
1,21–1,25	—	—	—	2	3	2	3	1	—	2	—	—
1,26–1,30	2	—	—	1	3	1	—	1	3	3	—	—
1,31–1,35	2	—	1	—	1	—	2	2	2	—	1	—
1,36–1,40	2	—	1	2	2	2	1	1	—	—	—	—
1,41–1,45	3	1	2	—	2	—	—	1	3	1	2	—
1,46–1,50	—	1	1	2	1	—	—	2	2	—	—	1
1,51–1,55	1	1	—	—	1	—	1	—	—	—	—	1
1,56–1,60	2	1	2	2	1	1	1	—	2	—	1	2
1,61–1,65	—	1	1	1	—	1	—	—	—	3	1	—
1,66–1,70	—	1	2	—	1	—	—	—	2	—	1	1
1,71–1,75	1	2	—	1	1	1	2	—	—	—	1	—
1,76–1,80	1	2	2	—	—	—	—	—	—	—	1	—
1,81–1,85	—	—	1	3	—	2	—	—	—	2	—	1

Tableau XI (*suite*)

Ville de Padoue : Fréquence mensuelle des jours de pluie (1725-1930)

Valeurs de l'indice de dispersion Q_2	Nombbes des indices de dispersion ayant la valeur indiquée à la col. 1 pour le mois de											
	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
1	2	3	4	5	6	7	8	9	10	11	12	13
1,86–1,90	1	—	1	2	—	—	—	—	—	1	1	→3
1,91–1,95	—	1	1	—	—	—	—	—	—	1	3	—
1,96–2,00	1	2	1	—	—	—	—	1	—	—	—	3
2,01–2,05	1	1	—	—	—	—	—	—	—	1	1	—
2,06–2,10	—	—	—	—	—	—	—	—	1	—	—	2
2,11–2,15	—	—	1	—	—	—	—	—	—	—	—	1
2,16–2,20	2	1	—	—	—	—	—	—	—	1	2	—
2,21–2,25	—	2	—	—	—	—	—	—	—	—	—	2
2,26–2,30	—	2	—	—	—	—	—	—	—	—	—	—
2,31–2,35	—	1	—	—	—	—	—	—	—	—	—	—
2,36–2,40	—	—	—	—	—	—	—	—	—	—	—	—
2,41–2,45	—	—	—	—	—	—	—	—	—	—	—	—
2,46–2,50	—	—	—	—	—	—	—	—	—	—	—	1
2,51–2,55	—	—	—	—	—	—	—	—	—	—	—	—
2,56–2,60	—	—	—	—	—	—	—	—	—	—	—	—
2,61–2,65	—	—	—	—	—	—	—	—	—	—	—	—
2,66–2,70	—	—	—	—	—	—	—	—	—	—	1	—
Totaux	20	20	20	20	20	20	20	20	20	20	20	20
Moyennes des valeurs de l'indice de dispersion Q_2'	1,597	1,889	1,595	1,477	1,328	1,226	1,219	1,203	1,386	1,483	1,678	1,836
Fréquences moyennes des jours de pluie	0,25	0,24	0,28	0,33	0,35	0,33	0,25	0,23	0,26	0,31	0,31	0,27

distribution des valeurs de Q'_2 se retrécit. D'autre part, les valeurs de l'indice de dispersion tendent à augmenter avec le nombre des années considérées dans chaque série, en passant d'une moyenne de 1,14, pour les séries à deux termes, à la valeur presque double de 2,17 pour la série de 210 termes.

Pour obtenir les valeurs de Q'_2, nous avons d'abord calculé les valeurs de Q'_3 auxquelles nous avons ensuite appliqué le coefficient de correction $\sqrt{1 - \frac{m}{n}}$ (voir page 7). Le coefficient n'est pas négligeable dans notre cas, car la valeur de m est en moyenne $= 104$, tandis que $n = 365$. La valeur moyenne du dit coefficient est à peu près $= 0{,}85$. Le coefficient de correction peut être négligé lorsque la valeur de $\frac{m}{n}$ ne dépasse pas 1 %, car, dans ce cas, le coefficient de correction n'a d'influence que sur le quatrième chiffre significatif. C'est le cas, ainsi que nous le verrons, pour les jours de neige et pour les jours de grêle dans certains pays.

Il est entendu que, dans ces cas, on applique la formule (3.1) seulement en vue d'obtenir une mesure approchée de la formule (2.1). La valeur de Q'_3 ne donne pas la mesure de la dispersion des nombres absolus des chutes de neige ou des averses de grêle, auxquels la formule (3.1) serait proprement applicable, car il y a des chutes de neige qui se prolongent pour plusieurs jours, tandis que, dans une même journée, on peut avoir plusieurs chutes de neige ou plusieurs averses de grêle.

Nous avons calculé aussi les valeurs de Q'_2 pour les jours de pluie des différents mois à Padoue pendant deux siècles, en partageant la période 1731-1930 en vingt séries de dix ans chacune. Le tableau XI donne la classification des valeurs de Q'_2 ainsi que les moyennes respectives desdites valeurs et de la fréquence des jours de pluie dans les différents mois.

Les fréquences présentent deux maxima : un au printemps (avril-juin) l'autre en automne (octobre-novembre), tandis que les valeurs moyennes de Q'_2 présentent un maximum dans l'hiver et un minimum dans l'eté Il n'y a donc pas une relation nette entre les variations de deux séries de moyennes : le coefficient de corrélation a la valeur assez faible de $r = -0{,}20$.

Des données intéressantes sur la fréquence des jours de pluie ont été publiées pour la ville de Pise par M. A. Rastrelli (1). Elles couvrent une période de 55 ans, de 1878 à 1932, et permettent d'examiner la dispersion de la fréquence des jours de pluie selon les différents mois, ainsi que selon les divers jours du même mois.

L'indice de dispersion Q'_2 montre que la fréquence des jours de pluie varie systématiquement selon les divers mois d'une façon très accentuée, sa valeur étant = 9,48 (2).

Au contraire, le jour du mois n'exerce aucune influence systématique sur la fréquence de la pluie. En effet, en considérant la fréquence de la pluie dans les jours successifs du mois (3), on obtient un indice de dispersion Q'_2 = 1,02.

Si on considère séparément la première décade, la deuxième décade et les 8 jours successifs, les valeurs de Q'_2, qui mesurent la dispersion selon les jours desdites périodes, sont respectivement = 0,80, 0,71 et 1,19 et montrent que la position du jour n'exerce aucune influence systématique sur la fréquence de la pluie.

Ces conclusions sont confirmées par les indices de dispersion calculés séparément pour chaque mois. Les valeurs obtenues sont exposées ou tableau XII.

Trois fois sur douze, pour le mois entier, et 19 fois sur 36, pour les périodes décadales (4), l'indice de dispersion reste inférieur à l'unité. Ces résultats suggèrent que la fréquence de la pluie ne varie presque pas systématiquement selon les jours successifs du même mois et ne varie pas du tout systématiquement selon les jours successifs de la même décade.

Etant établi que les variations de Q'_2 ont un caractère accidentel, il est naturel qu'elles ne peuvent avoir aucune relation systé-

(1) *Sulla frequenza delle pioggie a Pisa*, « Bollettino della Facoltà di Agraria della Regia Università di Pisa », 1937.

(2) Pour éliminer l'influence de la différente longueur des mois, on a calculé le nombre des jours de pluie comme si les mois avaient tous la longueur de trente jours.

(3) Pour ce calcul on a considéré les premiers 28 jours de tous les mois, en négligeant les autres.

(4) Les deux premières périodes du mois sont exactement de dix jours pour tous les mois ; la troisième période est de 11, 10 ou 8 jours selon la longueur du mois.

Tableau XII

Fréquence des jours de pluie dans les divers mois et dans les décades successives à Pise dans la période 1878-1932

MOIS	Valeurs de l'indice de dispersion Q'_2				Nombre moyen des jours de pluie sur un total de 55			
					dans la			dans tout le mois
	première décade	deuxième décade	troisième décade	mois entier	première décade	deuxième décade	troisième décade	
1	2	3	4	5	6	7	8	9
Janvier	0,64	1,48	0,97	1,11	19,60	17,20	16,73	17,74
Février	0,88	0,99	1,06	0,99	17,—	18,40	19,88	18,32
Mars	1,05	1,16	0,97	1,26	21,10	17,20	23,55	20,71
Avril	0,99	1,18	0,81	1,00	22,—	19,90	20,50	20,80
Mai	0,93	1,11	1,80	1,45	21,—	16,—	16,45	17,77
Juin	1,09	0,76	1,26	1,27	12,10	15,50	9,30	12,30
Juillet	0,80	0,91	1,40	1,12	8,70	6,30	6,45	7,13
Août	1,11	0,55	0,57	0,96	7,—	7,70	10,73	8,55
Septembre	0,51	1,44	0,62	1,12	12,60	15,50	17,60	15,23
Octobre	0,82	0,58	0,67	0,76	21,60	20,90	24,—	22,23
Novembre	1,03	0,65	1,22	1,07	26,80	22,80	23,60	24,40
Décembre	1,02	1,09	1,13	1,11	24,40	21,60	21,91	22,61

matique avec les variations de la fréquence de la pluie, qui montre deux maximums, un qui va de la dernière décade de février à la première décade de mai et l'autre, plus prononcé, dans les mois d'octobre, novembre et décembre.

En effet, le calcul des coefficients de corrélation entre les indices de dispersion (donnés aux colonnes 2, 3, 4, 5) et les fréquences respectives (données aux colonnes 6, 7, 8, 9) porte à des valeurs de r tantôt positives tantôt negatives (respectivement : + 0,14 ; + 0,12 ; — 0,23 ; — 0,14), qui confirment qu'il n'y a pas de corrélation systématique entre les deux phénomènes.

Si, enfin, on considère la fréquence de la pluie dans les jours classés en même temps selon le mois et selon leur succession dans le mois, on obtient — ainsi qu'il était à prévoir — des indices de dispersion supernormale, car dans ce cas les variations systématiques d'un mois à l'autre s'ajoutent aux variations purement accidentelles d'un jour à l'autre. Les valeurs de l'indice Q'_2 sont : 1,95 pour les jours de la première décade, 1,74 pour ceux de la deuxième décade ; 1,95 pour les 8 jours suivants et 1,88 pour tous les 28 jours considérés.

Sixième application : Jours de neige.

Nous avons deux longues séries de données sur les jours de neige à Rome et à Padoue.

Les données pour Rome ont été tirées de l'article de M. G. Roncali, *La neve a Roma*, « Rivista di Meteorologia », vol. II, fasc. 1-2, 1940 ; elles vont de l'année 1776 à l'année 1940 ; celles de Padoue vont de l'année 1725 à l'année 1921.

Les tableaux XIII et XIV donnent les valeurs de Q'_2 pour des périodes d'à-peu-près 20, 40, 80 années et pour une période encore plus longue couvrant 160 années pour Rome et 197 années pour Padoue.

Les valeurs de Q'_2 sont du même ordre de grandeur dans les deux villes, tournant autour de la valeur 1,5, mais plus fai-

TABLEAU XIII

Fréquence annuelle des jours de neige à Rome (Période 1776-1940)

PÉRIODES	VALEURS DE L'INDICE DE DISPERSION Q'_2				FRÉQUENCE MOYENNE DES JOURS DE NEIGE			
1	2	3	4	5	6	7	8	9
1776-1782	0,778	1,230			1,9	1,77		
1783-1801	1,400				1,7			
			1,283				1,77	
1802-1820	1,360	1,333			2,3	1,77		
1821-1840	1,158				1,25			
				1,407				1,72
1841-1861	1,342	1,338			1,3	1,46		
1862-1882	1,370				1,6			
			1,511				1,68	
1882-1901	1,624	1,637			2,8	1,9		
1902-1921	1,171				1,0			
1922-1940	1,560				1,7			

Tableau XIV

Fréquence annuelle des jours de neige à Padoue (Période 1725-1921)

PÉRIODES	Valeurs de l'indice de dispersion Q'_3				Fréquence moyenne des jours de neige			
1	2	3	4	5	6	7	8	9
1725-1744	1,44	1,57			4,0	5,2		
1745-1763	1,53				6,4			
1764-1782	1,46	1,81			7,2	9,1		
1783-1801	1,88		1,77		10,9		8,3	
1802-1820	1,54	1,70		1,76	7,5	7,5		6,5
1821-1840	1,88				7,6			
1841-1861	1,78	1,66			6,4	5,3		
1862-1881	1,32		1,55		4,3		5,3	
1882-1901	1,43	1,44			6,1	5,4		
1902-1921	1,42				4,6			

bles pour Rome (où les jours de neige sont moins fréquents) que pour Padoue : pour Padoue, elles dépassent en géneral ce chiffre, tandis que pour Rome elles en restent en général au-dessous.

Aussi bien à Rome qu'à Padoue, il n'y a pas de tendance dans les valeurs de Q'_2 à varier avec le temps. Aussi bien à Rome qu'à Padoue, la valeur de Q'_2 tend à augmenter lorsque le nombre des termes de la série augmente, mais l'augmentation est moindre que celle qui a lieu pour le nombre des jours de pluie, ce qui suggère que les variations de fond à travers le temps sont moins importantes pour la fréquence des jours de neige que pour la fréquence des jours de pluie.

À Padoue, et plus encore à Rome, les jours de neige sont rares : en moyenne 6,5 jours par an à Padoue ; pas même deux jours à Rome. Le coefficient de correction $\sqrt{1-\frac{m}{n}}$ a toujours été calculé ; pour Padoue, il influence légèrement les valeurs de l'indice Q'_2 arrêtées au troisième chiffre significatif ; pour Rome, il ne les influence pas.

Septième application : Jours de grêle.

Nous avons une longue série de données pour Padoue, qui couvre la période de 168 années de 1764 à 1921, donnant pour chaque année le nombre des jours où il y a eu de la grêle sur la ville. Ce nombre est assez petit (en moyenne inférieur à 3,4) de sorte que le coefficient $\sqrt{1-\frac{m}{n}}$ peut être négligé et la valeur de Q'_3 peut être calculée à la place de Q'_2.

Les valeurs de Q'_3 sont contenues dans le tableau XV. Leurs variations dans le temps sont assez sensibles, mais sans une tendance claire à augmenter ou à diminuer. Il y a au contraire une légère tendance de Q'_3 à augmenter avec l'augmentation du nombre des termes de la série. Ces résultats sont tout à fait analogues à ceux que l'on a obtenus pour les jours de neige concernant la même ville (Padoue), ainsi que pour ceux concernant Rome.

La valeur de l'indice de dispersion est toutefois sensiblement plus basse pour les jours de grêle que pour les jours de neige.

Des autres données [1] que nous avons élaborées pour quelques « Länder » allemands (Bade, Bavière, Wurtemberg) conduisent à des résultats différents. Il faut pourtant rappeler que leur signification est aussi différente.

Lorsque l'on dit que, dans la ville de Padoue, il y a eu quatre jours de grêle dans une année, cela signifie que, dans

---

[1] Les données sont tirées de deux thèses de doctorat présentées à des Universités allemandes : W. Rohrbeck, *Die Organisation der Hagelversicherung vornehmlich in Deutschland*, Friedrich-Wilhelms-Universität zu Berlin, 1909 ; W. Kleine, *Statistische Untersuchungen für die Hagelversicherung in Bayern in den Jahren 1884-1929*, Georg August-Universität zu Göttingen, 1933.

Tableau XV

Fréquence annuelle des jours de grêle à Padoue (Période 1764-1921)

PÉRIODES	Valeurs de l'indice de dispersion Q^3				Fréquence moyenne des jours de grêle			
1	2	3	4	5	6	7	8	9
1764-1782	0,89	0,93	1,17	1,23	4,2	4,2	4,0	3,3
1783-1801	0,98				4,2			
1802-1820	1,08	1,39			5,6	3,9		
1821-1840	1,15				2,3			
1841-1861	0,99	1,33	1,16		2,0	2,3	2,7	
1862-1881	1,54				2,6			
1882-1901	1,08	0,99			3,2	3,1		
1902-1921	0,91				3,0			

ces quatre jours, la grêle est tombée pratiquement sur toute l'étendue de la ville ; mais, lorsque l'on dit qu'en 1871 dans le pays de Bade il y a eu 120 jours de grêle, cela ne signifie pas que sur toute la superficie du pays la grêle soit tombée presque un jour sur trois. En 120 jours, il y a eu des averses de grêle dans le pays de Bade, mais chacune de ces averses de grêle n'en a évidemment affecté qu'une petite partie. Plus la surface du pays est étendue, plus, à égalité d'autres circonstances, le nombre des jours de grêle par année doit être élevé. D'autres circonstances pourtant peuvent exercer aussi de l'influence. En Bavière, le nombre moyen des jours de grêle par année (période 1890-929) a dépassé 86, tandis que dans le pays de Bade, pendant la même période, il n'a été que de 40, ce qui s'accorde avec la plus grande étendue de la Bavière, mais dans le Wurtemberg, plus étendu que le pays de Bade, il n'a été, dans la période 1894-905, que

de 11, tandis que, dans la même période, il a été de 46 dans ce dernier pays.

Nous pouvons conclure que la signification des chiffres donnant les jours de grêle non seulement est différente lorsqu'il s'agit d'un Etat ou d'un territoire étendu, d'un côté, et d'une ville ou d'une zone restreinte, de l'autre, mais aussi qu'elle est différente d'un Etat (ou territoire étendu) à un autre.

Il n'est pas surprenant, par conséquent, que les valeurs de Q_2' soient différentes pour des pays divers.

Pour le pays de Bade, on a une série assez longue, et la valeur de Q_2' pour la période totale (1868-936) est bien supérieure à la moyenne des valeurs pour les trois séries partielles, ainsi qu'il ressort du petit tableau suivant :

	Q_2'		A	
1868-87.	2,98		62	
1888-07.	1,93	2,91	47	46
1908-36.	1,95		35	

Cela fait comprendre que les fréquences des jours de grêle varient systématiquement au cours du temps, ce qui est confirmé par les valeurs du nombre moyen (A) des jours de grêle par année, qui diminue fortement de la première à la dernière série partielle.

Pour la Bavière, les valeurs de Q_2' sont 1,24 pour la période 1890-909, 1,56 pour la période suivante 1910-29 ainsi que pour la série complète 1890-1929. L'indice de dispersion pour la série totale n'est pas sensiblement plus élevé que la moyenne des indices pour les séries partielles, la fréquence des jours de grêle restant pratiquement invariée dans les deux périodes (respectivement 84,5 et 88,5 jours de grêle par an).

Les indices de dispersion pour la Bavière sont donc nettement inférieurs à ceux qu'on a obtenu pour le Bade.

Enfin, pour le Wurtemberg (période 1894-905), l'indice de dispersion ($Q_2' = 0{,}78$) n'atteint pas même l'unité correspondant à une dispersion normale.

Pour la Bavière, on a aussi calculé la dispersion pour les divers mois dans chacune des deux périodes 1890-909 et 1910-1929. Les variations mensuelles des valeurs de Q_2' sont irrégu-

lières et ne présentent aucune analogie pour les deux périodes ; le coefficient de corrélation est de $r = -0{,}22$. Elles n'ont pas non plus de relation nette avec les variations des fréquences mensuelles du nombre des jours de grêle, qui présentent une allure très régulière avec un maximum en été et un minimum en hiver. Le coefficient de corrélation est $r = -0{,}53$ pour la première période et $r = -0{,}26$ pour la deuxième. C'est bien ici un cas dans lequel les valeurs du coefficient de corrélation, si elles n'étaient pas accompagnées de l'examen des chiffres, pourraient prêter à des conclusions erronées. Les valeurs mensuelles de l'indice de dispersion et du nombre moyen des jours de grêle sont contenues dans le tableau XVI.

TABLEAU XVI

Fréquences mensuelles des jours de grêle et leurs indices de dispersion en Bavière (Périodes 1890-909 et 1909-929)

MOIS	PÉRIODE 1890-09		PÉRIODE 1909-29	
	Fréquences des jours de grêle	Valeurs de Q_2'	Fréquences des jours de grêle	Valeurs de Q_3'
Janvier	0,026	1,39	0,039	1,41
Février	0,035	0,92	0,051	1,49
Mars	0,084	1,25	0,129	1,73
Avril	0,263	0,98	0,360	1,85
Mai	0,502	1,08	0,556	1,46
Juin	0,552	0,83	0,673	1,26
Juillet	0,550	0,85	0,492	1,84
Août	0,411	0,85	0,297	1,50
Septembre	0,217	1,34	0,152	1,10
Octobre	0,085	1,30	0,053	1,72
Novembre	0,025	0,91	0,058	2,00
Décembre	0,026	1,13	0,039	1,98

Huitième application : Cas de maladie.

Toutes les applications précédentes concernent des séries chronologiques ; voyons maintenant une application à une série territoriale. Il s'agit des cas de maladie pour lesquels des indemnités ont été liquidées en 1952 dans les diverses régions d'Italie à des inscrits à l'ENPDEP (Ente Nazionale di Previdenza per i Dipendenti da Enti di Diritto Pubblico). Les données dont je

dispose sont le résultat d'un échantillon, qui comprend 16,8 % des cas de maladie et que la dite organisation a eu l'amabilité de tirer expressément de ses documents en vue de cette application.

Le nombre des inscrits varie très fortement d'une région à l'autre, ainsi qu'il ressort du tableau XVII qui suit. Il faut donc appliquer la formule (2).

La valeur de Q_2 est = 5,00.

Tableau XVII

Nombre des cas des maladie pour lesquels des indemnités ont été liquidées en 1952 à des inscrits à l'E. N. P. D. E. P.

Régions	Nombres des inscrits	Cas de maladie
Piémont (y compris le val d'Aoste)	8.197	758
Lombardie	11.818	989
Trentin, Haut-Adige	1.892	207
Vénétie	8.152	798
Frioul, Vénétie Julienne	1.603	169
Ligurie	5.211	490
Emilie, Romagne	10.495	1.132
Toscane	9.880	1.012
Ombrie	1.724	141
Marches	2.977	255
Latium	34.018	3.417
Abruzzes et Molise	2.570	185
Campanie	9.599	1.172
Pouille	6.607	564
Basilicate	987	58
Calabre	2.940	197
Sicile	9.312	576
Sardaigne	2.235	97
Totaux	130.217	12.217

9. Arrivé à la fin de ce premier groupe d'applications, on est tenté de comparer les indices de dispersion des divers phénomènes étudiés.

Pour huit des neuf applications que nous avons faites, nous avons considéré des séries annuelles.

Le fait que nous avons constaté que, plus ou moins pour tous les phénomènes, la dispersion augmente avec le prolongement de la série d'années considérées, nous oblige à prendre en considération, dans nos comparaisons, des séries qui ne diffè-

rent pas trop au point de vue du nombre d'années qu'elles comprennent. Nous avons choisi les séries de 10-20 années, car nous avons calculé la dispersion de séries de cette longueur pour tous les phénomènes étudiés.

Les indices de dispersion présentent un niveau minimum pour les précipitations atmosphériques, pour lesquelles les indices, dans la plupart des cas, n'excèdent pas deux unités.

L'indice est le plus bas pour les jours de grêle. Si on limite l'observation à des villes ou à des zones restreintes, il dépasse à peine l'unité, tandis que, pour les jours de neige, il est de 1,2-1,5 et, pour les jours de pluie, de 1,5-1,7. Légèrement plus élevé est l'indice de dispersion pour les jours qui ont présenté une précipitation atmosphérique quelconque. Pour les villes de l'Italie, sa moyenne varie de 1,5 à 2 (1).

Plus élevés doivent naturellement être les indices concernant des Etats entiers ou des zones très étendues. Pour les jours de grêle, les indices concernant des « Länder » allemands présentent des valeurs qui vont de 1 à 3.

Les tremblements de terre présentent apparemment une dispersion plus élevée, les indices étant compris entre 2 et 3 ; mais la supériorité n'est probablement qu'apparente, car les données sur les tremblements de terre concernent l'Italie, c'est-à-dire tout un Etat. Il est possible que, pour des zones restreintes, l'indice descendrait au même niveau que nous avons observé pour les précipitations atmosphériques.

Le fait que l'indice de dispersion pour un Etat ou une zone territoriale très étendue est en général plus élevé que pour une ville ou une zone territoriale restreinte provient du fait que, entre les variations de la fréquence du phénomène dans des zones contiguës, il y a généralement une corrélation positive, de façon que la dispersion du phénomène pour la surface totale desdites zones se montre supérieure à la dispersion de chacune d'entre elles.

Le fait que l'indice de dispersion, pour une zone ou une ville ou un pays déterminés, tend à augmenter lorsque la série

---

(1) Ces données concernant les précipitations atmosphériques sont tirées de recherches en cours, qui seront publiée prochainement.

des années se prolonge, dépend des tendances systématiques qui se font valoir dans la fréquence annuelle du phénomène. L'existence de ces tendances tend à rendre la dispersion supernormale, mais il se peut que, pour de courts intervalles, ces tendances soient neutralisées par des tendances à la compensation. Lorsque, par exemple, nous calculons la dispersion de la fréquence de la pluie dans les jours successifs du même mois et plus encore de la même décade, nous trouvons des indices de dispersion normale, ce qui peut être attribué au fait que, lorsque, dans un jour ou dans quelques jours successifs, il y a eu de la pluie, il est moins probable que les jours suivants soient aussi pluvieux : cette tendance compensatrice peut bien neutraliser la tendance systématique qui existe, mais se manifeste seulement pour des intervalles plus longs pour lesquels la tendance compensatrice a moins d'importance.

Il n'est pas dit non plus, d'ailleurs, que, plus l'intervalle est long, plus élevée doive être la dispersion, car dans un intervalle plus long des compensations peuvent se produire qui dans l'intervalle plus court n'interviennent pas. La dispersion de la fréquence des jours de pluie selon les mois ou les saisons de l'année est, par exemple, beaucoup plus élevée que celle selon les années ; selon les années, pour des séries de 10-20 termes, l'indice de dispersion, ainsi que nous l'avons vu, est = 1.5-1,7 ; selon les mois, = 9-10.

Si nous quittons le domaine des phénomènes purement physiques, tels que les pluies, les chutes de neige, les averses de grêle, les tremblements de terre, pour passer à ceux dans lesquels l'homme aussi intervient avec sa volonté ou ses négligences, la dispersion devient plus élevée.

Pour les cas d'incendie assurés auprès des compagnies d'assurance des Cantons suisses d'Argovie et de Berne, l'indice de dispersion présente des valeurs qui varient entre 1,5 et 4.

Lorsque, enfin, on passe à des phénomènes dans lesquels le résultat dépend de la combinaison des circonstances physiques, de la conduite humaine et de la conduite d'autres espèces animales, la dispersion monte à un ordre de grandeur supérieur. Pour les oiseaux tués par les chasseurs dans les « valli » de la Vénétie, l'indice de dispersion est de 8-12 unités et, pour les

thons pêchés dans les madragues du Sud-Ouest de la Sardaigne, il atteint des valeurs, pour des séries de 20 termes, de 20-27 unités.

10. Des applications que nous avons faites restent exclues les grandeurs mesurables absolues (telles que : distances parcourues, espaces de temps passés, quantité de pluie tombée, montant des profits réalisés, valeurs des diamants découverts, etc.) ou rapportées à d'autres grandeurs mesurables (telles que : dommages causés par les incendies rapportés à la valeur des bâtiments exposés, dommages causés par la grêle rapportés à la valeur des récoltes pendantes).

L'impossibilité de ces applications dérive du fait que les formules ci-dessus amèneraient à des résultats différents selon l'unité de mesure adoptée.

Or, la théorie de la dispersion peut aussi être étendue, comme nous le verrons, à des grandeurs de ce type, mais pour cela elle a besoin d'une généralisation ultérieure, à laquelle nous allons procéder en introduisant le concept de *dispersion d'ordre supérieur*.

11. Dans les schémas de la dispersion dont il a été question jusqu'ici, une urne seulement était considérée : la dispersion ainsi mesurée peut être appelée *dispersion du premier ordre*. Par opposition, nous appellerons *dispersion d'un ordre supérieur* celle qui est mesurée sur la base de schémas impliquant le recours à plusieurs urnes.

Pour illustrer ce concept, nous prenons en considération les grandeurs mesurables dont nous avons donné des exemples ci-dessus : distances parcourues, espaces de temps passés, quantités d'eau tombée, valeurs des diamants découverts, montants des profits réalisés, etc.

Retenons aussi que, pour les fins de la statistique, qui s'occupe de phénomènes collectifs, ce qui intéresse ce sont des collections ou masses de ces grandeurs, par exemple : distances parcourues non pas par un seul train, mais par tous les trains qui ont fonctionné sur une ou plusieurs lignes durant un certain espace de temps ; durée non pas d'une pluie seulement, mais de toutes les pluies qui se sont produites au cours d'une

année ; analoguement, quantité de l'eau tombée au cours non pas d'une seule, mais de toutes les précipitations atmosphériques de l'année ; valeur non pas d'un diamant seulement, mais de tous les diamants trouvés dans un certain territoire pendant un espace de temps donné ; montant des profits réalisés non pas dans une seule entreprise, mais dans toute une masse d'entreprises. Nous nommerons en général *montant extensif* le total des grandeurs mesurables ou extensives qui rentrent dans le phénomène collectif. Les montants extensifs que l'on trouve dans la statistique peuvent donc toujours être figurés comme les produits de deux grandeurs, variables l'une et l'autre : nombre des grandeurs qui composent le phénomène collectif et moyenne arithmétique des intensités de ces grandeurs. Le montant extensif peut être défini *grandeur mesurable à composantes énumérables.*

12. Nous dirons qu'une certaine série de montants (par exemple : la série des montants des lots de la loterie sortis dans chaque année d'un siècle) varie accidentellement, c'est-à-dire qu'elle présente une *dispersion normale du deuxième ordre,* lorsqu'il n'est pas possible d'apercevoir un facteur systématique dans les variations des grandeurs énumérables et mesurables dont résulte ce montant.

La dispersion sera dite du deuxième ordre parce que deux urnes interviennent dans le schéma des probabilités qui lui correspond.

Dans le cas des lots de la loterie, la première urne — l'urne des grandeurs énumérables — contiendra des boules de deux couleurs différentes, disons blanc et noir, les boules blanches correspondant aux billets gagnants, les noires aux billets perdants, et en nombre proportionnel à celui des billets gagnants et perdants dans les tirages effectués pendant la période considérée. La deuxième urne — l'urne des grandeurs mesurables — contiendra autant de boules qu'il y a de lots, chaque boule portant la mention du montant correspondant au lot respectif.

Si $n_1, n_2, n_3 \ldots . n_{100}$ sont les nombres des fois que l'on a joué pendant la première, la $2^{ème}$, la $3^{ème}$.... la $100^{ème}$ année du siècle, nous extrairons, avec remise, de la première urne successivement, $n_1, n_2, n_3 \ldots . n_{100}$ boules, en attribuant les boules blan-

ches extraites (au nombre de $m_1, m_2, m_3, \ldots m_{100}$) aux années respectives. Elles représentent le nombre des lots qui seraient sortis par l'effet du hasard dans les dites années.

A chacun de ces lots nous associerons chaque fois une boule prise au hasard sans remise dans la deuxième urne. Nous obtiendrons ainsi $m_1 + m_2 + \ldots . m_{100} = M$ grandeurs qui donneront lieu aux 100 montants $T_1 + T_2 \ldots . + T_{100} = T$. La dispersion de la série de ces 100 montants constituera la dispersion théorique des montants des lots sortis dans les 100 années du siècle, laquelle, comparée à la dispersion effective, nous dira si celle-ci est normale, ou bien supérieure ou inférieure à la normale.

Dans ce cas, le nombre des lots annuels possibles est limité au nombre des billets joués. Le schéma peut donc être dit de la *dispersion du deuxième ordre sur la base des rapports de dérivation.*

Si, par contre, le nombre des cas réalisés dans chaque section n'avait pas une limite supérieure définie, il nous faudrait recourir au schéma de la *dispersion de deuxième ordre sur la base des rapports de composition.*

Ce serait le cas pour la quantité d'eau tombée au cours des précipitations atmosphériques des différentes années d'un siècle. Il n'y a pas, en effet, de limite définie aux précipitations atmosphériques d'une année.

L'urne des grandeurs énumérables devrait, dans ce cas, contenir M boules (soit un nombre égal à celui des précipitations atmosphériques ayant eu lieu dans le siècle), marquées des numéros de 1 à 100 (soit le nombre des années du siècle) avec une fréquence proportionnelle à l'espace de temps correspondant à chaque année, c'est-à-dire au nombre de jours qu'elle contient (365 pour les années normales, 366 pour les années bissextiles). De cette urne on extraira, l'une après l'autre, avec remise, M boules en les attribuant à l'année marquée sur chacune d'elles. A chaque boule ainsi extraite de l'urne des grandeurs énumérables sera associée une boule extraite sans remise de l'urne des grandeurs mesurables, qui contiendra M boules correspondant aux précipitations atmosphériques et portant la mention de la quantité d'eau tombée au cours de chacune d'elles. On obtiendra ainsi une série composée de 100 montants, dont la dispersion pourra être comparée à la dispersion effective des

quantités de l'eau tombée au cours de chaque précipitation atmosphérique dans chaque année du siècle.

13. Considérons d'abord le cas plus simple où les sections sont équivalentes, dans le sens qu'il y a la même probabilité qu'une grandeur entre dans l'une ou l'autre d'entre elles (comme il arriverait dans l'exemple ci-dessus, si toutes les années avaient le même nombre de jours).

La somme des carrés des écarts des montants des s sections considérées : $T_1, T_2, \ldots T_s$ (en général T_k) de leur moyenne T/s, où $T = \sum_{k=1}^{s} T_k$, sera

$$\sum_{k=1}^{s} T_k^2 - T^2/s$$

La valeur théorique de cette expression, sur la base du schéma des rapports de composition, est [1]

$$\sum_{k=1}^{s} \sum_{i=1}^{m_k} a_{ik}^2 - T^2/Ms$$

où a_{ik} indique l'intensité du phénomène dans le $i^{ème}$ cas ($i = 1,2,\ldots m_k$) de la section $k^{ième}$.

On en tire l'indice de dispersion

$$Q_4 = \sqrt{\frac{\sum_{k=1}^{s} T_k^2 - T^2/s}{\sum_{k=1}^{s} \sum_{s=1}^{m_k} a_{ik}^2 - T^2/Ms}} \qquad (4)$$

Pour a_{ik} = constante, la formule (4) se réduit à la formule (3,1).

Sur la base du schéma des rapports de dérivation, la valeur de l'expression $\sum_{k=1}^{s} T_k^2 - T^2/s$ est par contre

$$\frac{n\,(s-1)}{ns-1}\left(\sum_{k=1}^{s} \sum_{i=1}^{m_k} a_{ik}^2 - T^2/ns\right)$$

---

[1] V. l'article cité *Estensione della teoria della dispersione e della connessione*, etc., § 8.

De là on déduit l'indice de dispersion

$$Q_5 = \sqrt{\frac{\sum_{k=1}^{s} T_k^2 - T^2/s}{\frac{n(s-1)}{ns-1}\left(\sum_{k=1}^{s}\sum_{i=1}^{m_k} a_{ik}^2 - T^2/ns\right)}} \qquad (5)$$

qui, pour a_{ik} constante, se réduit à la formule (2,1).

Il est évident que les valeurs des indices de dispersion qu'on tire des formules (4) et (5) sont indépendantes de l'unité de mesure.

14. Passons au cas plus général, dans lequel les s sections ne sont pas équivalentes, mais présentent une probabilité différente qu'une grandeur rentre dans l'une ou dans l'autre d'entre elles.

C'est, par exemple, le cas lorsque nous voulons examiner la dispersion des montants des dommages causés par les tremblements de terre dans diverses circonscriptions territoriales ayant une étendue différente, ou bien la dispersion des montants des indemnités à payer aux assurés qui ont été malades dans les diverses circonscriptions administratives qui comprennent des nombres différents d'assurés.

Dans ce deuxième exemple, nous devons appliquer le schéma des rapports de dérivation, car le nombre des assurés qui ont été malades présente une limite définie par le nombre des assurés ; dans le premier exemple, nous devons au contraire appliquer le schéma des rapports de composition, car on ne peut pas fixer une limite au nombre des tremblements de terre dans une circonscription territoriale.

L'indice de dispersion du deuxième ordre sur la base des rapports de composition est donné par la formule

$$Q_6 = \sqrt{\frac{\sum_{k=1}^{s}\left(\frac{T_k}{T} - \frac{n_k}{N}\right)^2}{\frac{\sum_{k=1}^{s}\sum_{i=1}^{m_k} a_{ik}^2}{T^2} - \frac{\sum_{k=1}^{s} n_k^2}{MN^2}}} \qquad (6)$$

et l'indice de dispersion du deuxième ordre sur la base des rapports de dérivation par la formule :

$$Q_7 = \sqrt{\frac{\sum_{k=1}^{s} \left(\frac{T_k}{T} - \frac{n_k}{N} \right)^2}{\frac{N(s-1)}{(N-1)s} \left(\frac{\sum_{k=1}^{s} \sum_{s=1} a_{ik}^2}{T^2} - \frac{1}{N} \right)}} \qquad (7)$$

15. Il importe d'observer qu'une même série de montants peut être scindée de plusieurs manières en ses composantes énumérables et mesurables. Par exemple, la quantité de l'eau tombée au cours des précipitations atmosphériques de l'année peut être considerée comme le produit du nombre des précipitations par la quantité moyenne de l'eau tombée au cours de chaque précipitation, ou bien comme le produit du nombre des jours où se sont produites des précipitations atmosphériques par la quantité moyenne de l'eau tombée dans chacun d'eux.

Dans le premier cas, on applique le schéma des rapports de composition, puisqu'il n'y a pas de limite définie au nombre des précipitations qui peuvent avoir lieu dans une année ; dans le second cas, on applique le schéma des rapports de dérivation, puisque le nombre des jours où des précipitations se produisent a une limite définie par le nombre des jours de l'année.

Or, il n'est pas dit que l'application des deux schémas conduise au même résultat. Effectivement, le concept de dispersion normale dans un cas est différent du concept de dispersion normale dans l'autre. Dans le premier cas, il faut que les nombres des précipitations se distribuent accidentellement au cours de l'année ; dans le second, ce sont les nombres des jours avec précipitation qui doivent être ainsi distribués ; ce n'est pas la même chose, car il peut arriver que le nombre des jours où des précipitations se produisent varie accidentellement, mais non pas celui des précipitations, du moment que, dans un même jour, des précipitations peuvent se produire en grand ou en petit nombre et il peut aussi arriver qu'une seule précipitation se produise.

16. Dans quelques cas, la même série de grandeurs énumérables peut être traitée sur la base du schéma de la dispersion

du premier ordre ou sur la base de celui de la dispersion d'ordre supérieur.

Prenons l'exemple du nombre des secousses sismiques de terre ayant eu lieu dans diverses années. Elles constituent des grandeurs énumérables dont il est possible d'étudier la dispersion de premier ordre sur la base des rapports de composition.

Mais nous pouvons aussi scinder le nombre de ces secousses en deux composantes : le nombre des tremblements de terre et la moyenne des secousses pour chaque tremblement. Nous pourrons alors considérer ces grandeurs comme des *grandeurs énumérables à composantes énumérables.*

Nous donnerons à ces grandeurs la dénomination de *montants intensifs.*

Nous pourrons étudier leur dispersion du deuxième ordre sur la base du schéma des rapports de composition et appliquer la formule (4).

Evidemment, les résultats auxquels on arrive sur la base de ces deux schémas sont différents. Voyons la relation qui existe entre les indices de dispersion que l'on obtient au moyen des deux schémas.

17. La formule (4) peut être mise sous la forme

$$Q_4 = \sqrt{\frac{\sum_{k=1}^{s} T_k^2 - T^2/s}{(A\ T/s)\ (s\ B^2/A^2 - 1)}} \qquad (4 \text{ bis})$$

où $A = \frac{T}{M}$ indique la moyenne arithmétique et B la moyenne quadratique des M valeurs a_{ik}.

Selon le schéma de la dispersion de premier ordre, on devrait appliquer la formule (3.1), qui dans ce cas prendrait la forme

$$Q'_3 = \sqrt{\frac{\sum_{k=1}^{s} T_k^2 - T^2/s}{(T/s)\ (s-1)}} \qquad (3.1 \text{ bis})$$

D'où

$$Q_3' = Q_4 \sqrt{A \frac{s\, B^2/A^2 - 1}{s - 1}} \qquad (8)$$

Le coefficient qui figure dans le deuxième membre de cette formule est toujours positif, A étant toujours positif et B (moyenne quadratique) étant toujours supérieur à A (moyenne arithmétique).

L'indice de dispersion du premier ordre Q_3' est donc toujours supérieur à l'indice de dispersion correspondant du deuxième ordre Q_4 pour un coefficient qui est toujours supérieur à la racine carrée de la moyenne arithmétique des valeurs a_{ik}.

18. Dans les paragraphes qui précèdent, nous avons étudié les applications de la théorie de la dispersion aux nombres absolus (m_k) — tels que le nombre des tremblements de terre ou le nombre des cas d'incendie — ainsi qu'aux montants extensifs ou intensifs (T_k) — tels que le nombre des secousses sismiques ou les dommages des incendies.

Considérons maintenant les applications aux moyennes $\frac{T_k}{m_k} = A_k$.

Dans les schémas du deuxième ordre qui s'appliquent aux montants, nous avons supposé qu'à chaque tirage, une des M valeurs a_{ik} était choisie au hasard, sans remise, de l'ensemble des M valeurs (tirage de l'urne B) et attribué à celle des s sections qui avait été choisie au hasard avec remise (tirage de l'urne A).

On aura donc, pour les m_k valeurs attribuées à la section k,

$$E\left\{\frac{1}{m_k}\sum_{i=1}^{m_k}(a_{ik} - A)^2\right\} = \frac{1}{M}\sum_{k=1}^{s}\sum_{i=1}^{m_k}(a_{ik} - A)^2$$

(où E (.....) est l'espérance mathématique de.....) et

$$E\left\{\frac{1}{m_k}\sum_{i=1}^{m_k}(a_{ik} - A_k)^2\right\} = \frac{m_k - 1}{m_k}\,\frac{M}{M-1}\,\frac{1}{M}\sum_{k=1}^{s}\sum_{i=1}^{m_k}(a_{ik} - A)^2$$

Du fait que, par une propriété bien connue de la moyenne arithmétique,

$$\frac{1}{m_k}\sum_{i=1}^{m_k}(a_{ik}-A)^2=(A_k-A)^2+\frac{1}{m_k}\sum_{k=1}^{s}(a_{ik}-A_k)^2$$

on obtient

$$E\{(A_k-A)^2\}=\frac{1}{M}\sum_{k=1}^{s}\sum_{i=1}^{m_k}(a_{ik}-A)^2\left(1-\frac{(m_k-1)\,M}{m_k\,(M-1)}\right)=$$

$$=\frac{M-m_k}{m_k\,(M-1)}\,\frac{1}{M}\sum_{i=1}^{m_k}(a_{ik}-A)^2$$

Si, ainsi qu'il paraît naturel, on attribue à chaque carré de l'écart (A_k-A) un poids proportionnel au nombre m_k des éléments auxquels la moyenne A_k se rapporte, on obtient, pour l'indice de dispersion des moyennes,

$$Q''_A=\sqrt{\frac{(M-1)\sum_{k=1}^{s}m_k\,(A_k-A)^2}{(s-1)\left(\sum_{k=1}^{s}\sum_{i=1}^{m_k}a_{ik}^2-M\,A^2\right)}} \tag{9}$$

Si, au contraire, on attribuait à toutes les s valeurs de $(A_k-A)^2$ le même poids, l'indice de dispersion prendrait la forme

$$Q'_A=\sqrt{\frac{(M-1)\sum_{k=1}^{s}(A_k-A)^2}{s\left(\frac{1}{H}-\frac{1}{M}\right)\left(\sum_{k=1}^{s}\sum_{i=1}^{m_k}a_{ik}^2-M\,A^2\right)}} \tag{10}$$

où $H=s/\sum_{k=1}^{s}\frac{1}{m_k}$ est la moyenne harmonique des valeurs m_k, et, enfin, si on attribuait à chaque valeur $(A_k-A)^2$ un poids proportionnel à m_k^2, on parviendrait à l'expression

$$Q'''_A=\sqrt{\frac{(M-1)\sum_{k=1}^{s}m_k^2\,(A_k-A)^2}{(M-K)\left(\sum_{k=1}^{s}\sum_{i=1}^{m_k}a_{ik}^2-M\,A^2\right)}} \tag{11}$$

où $K=\sum_{k=1}^{s}m_k^2/\sum_{k=1}^{s}m_k$ est la moyenne antiharmonique des valeurs m_k.

Or, si nous revenons à la formule (1), nous trouvons qu'elle est identique à la formule (9), les symboles m_k et M ayant dans celle-ci la même signification que n_k et N dans celle-là.

Ce résultat pourra peut-être surprendre de prime abord, car la formule (1) donne un indice de dispersion du premier ordre, tandis qu'ici nous sommes dans le domaine de la dispersion du deuxième ordre. En réfléchissant, le résultat apparaît pourtant tout à fait naturel.

En effet, la démonstration qui précède, et qui a conduit aux formules (9), (10), (11), est valable pour des valeurs de m_k quelconques.

C'est-à-dire que, pour l'indice de dispersion des moyennes, les valeurs de m_k qui résultent des tirages de l'urne A n'ont pas d'importance ; autrement dit, l'urne A n'entre pas en question pour l'indice de dispersion des moyennes. Les indices de dispersion des moyennes Q_A sont des indices de dispersion du premier ordre, basés sur les tirages de l'urne B.

Au contraire, pour l'indice de dispersion Q'_3, donné par la formule (3.1), concernant les fréquences m_k, l'urne B n'entre pas en question : Q'_3 est un indice de dispersion du premier ordre basé sur les tirages de l'urne A.

Les urnes A et B sont, enfin, nécessaires, toutes les deux, pour déterminer l'indice de dispersion Q_4, concernant les montants, donné par la formule (4), qui est un indice de dispersion du deuxième ordre.

Cela suggère que l'indice des montants Q_4 doit indiquer une dispersion normale quand et l'indice de fréquence Q'_3 et l'indice des moyennes Q_A indiquent une dispersion normale. Ces deux conditions pourtant ne sont pas suffisantes, ainsi que nous allons le voir.

19. La formule (4) peut être mise sous la forme suivante

$$Q_4 = \sqrt{\frac{\sum_{k=1}^{s} (m_k A_k - m_k A)^2 + A^2 \sum_{k=1}^{s} (m_k - m)^2 + 2 \sum_{k=1}^{s} m_k (A_k - A) A (m_k - m)}{\sum_{k=1}^{s} \sum_{i=1}^{m_k} a_{ik}^2 - M A^2 + A^2 (s-1) m}} \qquad (4.1)$$

Si l'on indique par $r_{A_k m_k}$ le coefficient de corrélation (suivant la formule de Bravais) entre les A_k et les m_k

$$r_{A_k m_k} = \frac{\sum_{k=1}^{s} m_k (A_k - A)(m_k - m)}{\sqrt{\sum_{k=1}^{s} (m_k - m)^2 \sum_{k=1}^{s} m_k^2 (A_k - A)^2}}$$

dans lequel on attribue à chaque écart $(A_k - A)$ un poids m_k et par conséquent à son carré un poids m_k^2, et l'on fait

$$R' = A^2 (s - 1)\, m \qquad R'' = \sum_{k=1}^{s} \sum_{i=1}^{m_k} a_{ik}^2 - M A^2$$

$$R''' = \sqrt{\sum_{k=1}^{s} (m_k - m)^2 \cdot \sum_{k=1}^{s} m_k^2 (A_k - A)^2},$$

la formule (4.1) — ayant recours aux formules (4) et (11) — devient

$$Q_4 = \sqrt{\frac{M - K}{M - 1} \frac{R'' (Q_A''')^2 + R' (Q_3')^2 + 2\, r_{A_k m_k} R''' A}{R'' + R'}} \qquad (4.2)$$

Cette formule nous montre que, en négligeant le facteur $(M - K)/(M - 1)$, Q_4 devient $= 1$ lorsqu'on a en même temps $Q_3' = 1$, $Q_A''' = 1$ et $r_{A_k m_k} = 0$. C'est-à-dire qu'en négligeant ledit facteur, la dispersion des montants apparaît normale lorsque sont normales la dispersion des fréquences ainsi que la dispersion des moyennes, et en même temps la corrélation est nulle entre moyennes et fréquences.

Si la corrélation entre moyennes et fréquences est nulle, l'indice Q_4 de dispersion des montants est donné par une moyenne pondérée de l'indice Q_3' de dispersion des fréquences et de l'indice Q_A''' de dispersion des moyennes. Si la corrélation n'est pas nulle, l'indice Q_4 de dispersion des montants est supérieur ou inférieur à ladite moyenne pondérée selon que la corrélation est positive ou négative.

20. De tout ce que nous avons dit aux paragraphes précédents 14-19 il apparaît clair que les conclusions sur la dispersion normale, supernormale ou sousnormale d'une série n'acquièrent un sens précis que lorsque le schéma est précisé sur la base duquel on juge de la dispersion.

21. Nous sommes maintenant en état d'appliquer les formules pour la mesure de la dispersion d'un ordre supérieur.

I^ere application : *Secousses sismiques.*

Commençons par des montants intensifs, dont les secousses sismiques fournissent un exemple.

Celles-ci peuvent, ainsi que nous l'avons remarqué, être traitées aussi d'après le schéma de la dispersion du premier ordre. C'est-à-dire que nous pouvons comparer la variabilité que les nombres de secousses présentent en réalité d'une période à l'autre, ou d'une zone territoriale à l'autre, avec la variabilité théorique qu'ils présenteraient si les secousses étaient indépendantes l'une de l'autre et la probabilité d'une secousse ne variait pas dans le temps ou dans l'espace.

Nous avons les données pour les secousses qui se sont produites en Italie dans chacune des quarante années qui vont de 1891 à 1930 (1).

Voici les valeurs de Q'_3 que l'on obtient.

1891-900	8,93		
		11,95	
1901-910	12,95		
			11,97
1911-920	11,60		
		11,24	
1921-930	10,64		

L'autre série, pour les 14 années 1917-30 (voir tableau IX), conduit à une valeur de $Q'_3 = 9{,}64$.

Ces résultats, basés sur des chiffres puisés aux mêmes sources que les chiffres concernant les tremblements de terre, sont compa-

(1) Voir l'article de A. Cavasino cité à la page 26, note (2).

rables aux indices de dispersion obtenus pour ceux-ci (voir troisième application, pages 26 et sq.).

Les valeurs de Q'_3 sont régulièrement plus élevées (environ triples) pour les secousses que pour les tremblements de terre.

Pour l'ensemble de quarante années, la même source fournit des chiffres aussi pour le nombre des secousses dans les douze mois de l'année, ainsi que dans les diverses heures du jour, groupées deux à deux en douze classes, pour chacun des douze mois et pour chacune des quatre saisons.

La valeur de Q'_3, pour la dispersion selon les mois de l'année, est = 2,38 ; pour la dispersion selon les quatre saisons = 1,32 [1].

On est frappé de la différence très marquée entre ces valeurs et celles — beaucoup plus élevées — concernant les années. Malheureusement nous n'avons pas les chiffres des tremblements de terre selon les mois et les saisons qui permettraient de calculer les indices de dispersion respectifs.

A remarquer aussi la différence sensible entre l'indice de dispersion pour les saisons et celui pour les mois. Cela suggère qu'il y a une tendance à la compensation entre le nombre des secousses des trois mois appartenant à la même saison, et c'est en effet ce qui arrive : dans chaque saison, un des trois mois présente un écart de la moyenne générale de l'année de signe contraire aux écarts des deux autres.

Pour la dispersion de la fréquence horaire des secousses dans chaque mois, on obtient les valeurs suivantes de Q'_3 :

Décembre	2,94	Mars	2,41	Juin	3,00	Septembre	3,07
Janvier	3,73	Avril	2,10	Juillet	1,71	Octobre	2,05
Février	2,78	Mai	2,77	Août.	2,14	Novembre	2,88
Moyennes	3,15		2,43		2,28		2,67

Les valeurs pour la distribution horaire dans les quatre saisons sont remarquablement supérieures aux moyennes ci-dessus.

Hiver	5,07	Printemps	3,76	Etée	3,67	Automne	4 34

[1] Le nombre des secousses a été calculé comme si tous les mois avaient 31 jours.

Et encore plus élevée ($= 8{,}12$) est la valeur de Q'_3 pour la distribution horaire des secousses de toute l'année.

Cela provient du fait qu'il y a un parallélisme assez marqué entre les variations horaires des secousses dans les différents mois, de façon que les variations se cumulent — et la dispersion de la sorte augmente — en groupant les mois et les saisons.

L'application aux secousses du schéma de la dispersion du premier ordre n'est pourtant pas très instructive. On sait à priori que les secousses ne sont pas indépendantes l'une de l'autre et que, par conséquent, les données ne peuvent pas se conformer au dit schéma. Il est donc à prévoir que les indices de dispersion du premier ordre présentent des valeurs supérieures à l'unité, mais on ne sait pas la signification à attribuer à ces valeurs.

Un schéma plus instructif est évidemment celui de la dispersion du deuxième ordre, d'après lequel on compare la variabilité effective des secousses avec celle qui se produirait si les tremblements de terre variaient d'une période à l'autre par le seul effet du hasard et le nombre des secousses pour chaque tremblement de terre variait aussi au hasard d'un tremblement à l'autre.

Pour appliquer ce schéma, il faut avoir à sa disposition des données plus détaillées que celles qui sont suffisantes pour appliquer le schéma du premier ordre ; il faut, en effet, connaître, pour chaque intervalle considéré, non seulement le nombre des secousses, mais aussi la classification des tremblements de terre selon les secousses qu'ils ont présentées.

Nous avons ces données pour les tremblements de terre qui se sont produits en Italie dans les quatorze années 1917-1930 (voir tableau IX).

Nous avons vu que l'indice de dispersion Q'_3 pour les tremblements de terre de ladite période est $= 3{,}20$ et que l'application du schéma de la dispersion du premier ordre aux secousses porte à une valeur de $Q'_3 = 9{,}64$. Or, l'application aux mêmes chiffres des secousses du schéma de la dispersion du deuxième ordre porte à une valeur de $Q'_4 = 3{,}55$. Le rapport entre Q'_3 et Q'_4 est $= 2{,}72$. Le nombre moyen des secousses pour chaque tremblement de terre est $= 2{,}78$, dont la racine carrée est $= 1{,}66$. On a $1{,}66 < 2{,}72$, ainsi qu'on pouvait le prévoir (cf. § 17, pages (50-51)

L'indice de dispersion Q'''_A pour les secousses moyennes pour chaque tremblement est = 2,48.

La valeur de Q'_4 pour les secousses (3,55) est donc légèrement plus élevée que la valeur de Q'_3 (3,20) pour les tremblements de terre et sensiblement plus élevée que la valeur de Q'''_A (2,48) pour les moyennes des secousses pour chaque tremblement de terre. Cela suggère que, dans les années dans lesquelles il y a plus de tremblements de terre, il y a aussi plus de secousses pour chaque tremblement de terre. L'examen des données confirme cette prévision Le coefficient de corrélation de Bravais est en effet $r = 0{,}70$.

II^eme application : Incendies

Pour les incendies, nous avons des données qui nous permettent de faire des applications du schéma du deuxième ordre à des montants intensifs (nombre des bâtiments endommagés) aussi bien qu'à des montants extensifs (montants des dommages).

A) *Nombre des bâtiments endommagés.*

Les données ont été publiées par l'Institut d'Assurance contre l'incendie de l'Argovie, dans la publication citée à la page. 27 Elles concernent 336 cas d'incendie grave, avec 1929 bâtiments brûlés ou autrement endommagés, pendant un siècle (1806-1905). Elles ont été groupées en périodes de dix années.

La dispersion pour les cas d'incendie grave donne — ainsi que nous l'avons vu (page 28) — une valeur de $Q_2 = 3{,}14$. L'application du schéma de dispersion du premier ordre aux bâtiments brûlés ou endommagés donne une valeur de $Q_3 = 5{,}60$. Sur sa signification on pourrait faire des remarques analogues à celles que nous avons faites au sujet des secousses des tremblements de terre. L'application du schéma du deuxième ordre porte à une valeur de $Q_7 = 1{,}54$.

Dans ce schéma on suppose que le nombre des incendies varie au hasard d'une période décennale à l'autre et que le nombre moyen des bâtiments endommagés pour chaque cas d'incendie varie aussi au hasard d'un incendie à l'autre. Dans notre exemple, on connaît le nombre des bâtiments exposés au risque

d'incendie — il s'agit donc de rapports de dérivation — et ce nombre change d'une période décennale à l'autre ; c'est donc la formule (7) que nous avons appliquée.

D'autre part, l'indice de dispersion pour le nombre moyen des bâtiments endommagés est $Q''_A = 2{,}75$.

La valeur de $Q_7 = 1{,}54$ est donc remarquablement inférieure à celle de $Q_2 = 3{,}14$ pour les cas d'incendie grave, de même qu'à celle de $Q''_A = 2{,}75$ pour le nombre moyen des bâtiments endommagés. Cela suggère qu'il y a une compensation entre le nombre des cas d'incendie grave et le nombre moyen des bâtiments endommagés par chaque incendie. L'examen des données le confirme. Le coefficient de corrélation se chiffre en effet à $r = -0{,}76$.

Tout cela pour les cas d'incendie grave.

Pour l'ensemble des incendies (4.400 dans les 100 années 1806-1905) on a aussi le nombre des bâtiments endommagés, mais il n'est pas possible de faire l'application du schéma de dispersion du deuxième ordre, car on ne connaît pas la classification des cas d'incendie selon le nombre des bâtiments endommagés. L'application du schéma du premier ordre porte à des valeurs de Q_2 (1806-55 = 4,89 ; 1856-905 = 3,07 ; 1806-905 = 4,10) sensiblement supérieures à celles (respectivement 2,92 ; 2,63 ; 3,13) obtenues pour les cas d'incendie.

Les résultats sont analogues pour les incendies concernant l'Institut d'Assurance du Canton de Berne. Le schéma de la dispersion du premier ordre, appliqué aux édifices endommagés, porte à des indices (1807-34 : 4,08 ; 1835-58 : 4,96 ; 1859-82 : 6,07 ; 1883-906 : 3,92) bien supérieurs à ceux que nous avions obtenus pour les cas d'incendie (respectivement : 1,36 ; 3,06 ; 3,91 ; 2,39). Ici aussi on n'a pas de données qui permettent d'appliquer le schéma du deuxième ordre.

B) *Dommages causés par les incendies.*

Dans ce cas, on ne peut pas penser à appliquer le schéma de la dispersion du premier ordre.

L'application du schéma du deuxième ordre peut être faite soit à la série des dommages causés par les incendies graves d'après les statistiques de l'Institut de l'Argovie, soit à celle des

Nombre des jours de grêle et superficie endomm

ANNÉES	Nombre des jours de grêle dans lesqu										
	au-dessous de 900	de 900 à 1800	de 1800 à 2700	de 2700 à 3600	de 3600 à 4500	de 4500 à 5400	de 5400 à 6300	de 6300 à 7200	de 7200 à 8100	de 8100 à 9000	de à
1908	34	1	2	—	—	2	—	—	—	—	
1909	36	—	—	—	—	—	1	—	—	—	
1910	44	7	3	2	1	—	—	—	1	—	
1911	33	2	—	2	—	—	—	—	—	—	
1912	44	4	2	1	3	—	—	—	—	—	
1913	35	4	1	—	—	—	—	—	—	—	
1914	32	4	3	1	2	—	—	—	—	—	
1915	30	4	1	—	—	—	—	—	—	—	
1916	33	1	1	—	—	1	—	—	—	—	
1917	43	2	2	—	1	—	—	—	—	—	
1918	26	2	—	—	—	—	—	—	—	—	
1919	30	1	1	—	—	—	1	—	—	—	
1920	27	1	1	—	—	—	—	—	—	—	
1921	28	3	—	2	—	1	—	1	—	—	
1922	21	6	2	—	—	—	2	1	—	—	
1923	9	—	—	—	—	—	—	—	—	—	
1924	25	1	1	—	—	1	—	—	1	—	
1925	23	4	—	—	2	—	—	—	—	1	
1926	24	5	2	—	—	1	—	—	—	—	
1927	30	5	1	1	—	1	1	—	—	—	
1928	19	2	1	—	—	1	—	1	1	—	
1929	24	6	2	—	1	—	1	—	—	—	
1930	46	2	1	1	1	—	—	1	—	—	
1931	40	2	2	—	1	—	—	—	—	—	
1932	24	3	—	—	—	—	—	—	—	—	
1933	23	2	2	—	1	—	—	—	—	—	
1934	24	6	—	—	1	—	—	—	—	—	
1935	12	1	—	—	—	—	—	—	—	—	
1936	13	4	2	—	1	1	—	—	—	—	
Totaux	832	85	33	10	15	9	6	4	3	1	

Pour calculer la superficie totale des diverses classes, on a admis que la superficie moyenne était, pour la pr
superficies de 1800 à 2700 ha. qu'elle était = 2300 ha.).

Tableau XVIII

s *chaque jour de grêle* (*Bade, 1908-36*)

ICIE ENDOMMAGÉE (EN HECTARES) A ÉTÉ LA SUIVANTE :											Nombre total des jours de grêle
)00 à 300	de 10800 à 11700	de 11700 à 12600	de 12600 à 13500	de 13500 à 14400	de 14400 à 15300	de 15300 à 16200	de 16200 à 17100	de 17100 à 18000	de 18000 à 18900	18900 et au-dessus	
—	—	—	—	—	1	—	—	—	—	—	40
1	—	—	—	—	—	—	—	—	—	—	38
—	—	—	—	—	—	—	—	—	—	—	58
—	—	—	—	—	—	—	—	—	—	—	38
—	—	—	—	—	—	—	—	—	—	—	54
—	—	—	—	—	—	—	—	—	—	—	40
—	—	—	—	—	—	—	—	—	—	—	42
—	—	—	—	—	—	—	—	—	—	—	35
—	—	—	—	—	—	—	—	—	—	—	36
—	—	—	—	—	—	—	—	—	—	—	48
—	—	—	—	—	—	—	—	—	—	—	28
—	—	—	—	—	—	—	—	—	—	—	33
—	—	—	—	—	—	—	—	—	—	—	29
—	—	—	—	—	—	—	—	—	—	—	35
—	—	—	—	—	—	—	—	—	—	—	32
—	—	—	—	—	1	—	—	—	—	—	10
—	—	—	—	—	—	—	—	—	—	—	29
—	—	—	—	—	—	—	—	—	—	—	30
—	—	—	—	—	—	—	—	—	1	—	33
—	—	—	—	—	—	—	—	—	—	—	39
—	—	—	—	—	—	—	—	—	—	—	25
—	—	—	—	—	1	—	—	—	—	—	35
—	1	—	—	—	—	—	1	—	—	—	54
—	—	—	—	—	—	—	—	—	—	—	45
1	—	—	—	—	—	—	—	—	—	—	28
—	—	—	—	—	—	—	—	—	—	—	28
—	—	—	—	—	—	—	—	—	—	—	31
—	—	—	—	—	—	—	—	—	—	—	13
—	—	—	—	—	—	—	—	—	—	—	21
2	1	—	—	—	3	—	1	—	1	—	1.007

e, 500 ha. et, pour les autres classes, égale à la demi-somme des limites de la classe (par exemple, pour la classe des

Nombres des jours de grêle et nombre des commu

ANNÉES	NOMBRE DES JOURS DE GRÊLE DANS LESQUELS LE NO																													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	27	28	29	30	31	
1908	10	5	6	2	1	1	1	2	2	—	—	1	—	—	—	—	1	—	1	—	—	—	—	—	—	1	—	—	—	1	
1909	20	4	3	3	2	1	—	1	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	1	—	—	—	—	—	
1910	15	5	7	6	1	—	1	3	2	1	2	1	1	1	1	2	—	—	1	—	1	1	—	—	—	—	1	—	2	—	
1911	13	2	5	3	3	1	1	1	—	2	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1	—	1	—	1	
1912	17	11	3	3	4	1	2	1	—	—	—	—	—	—	—	—	—	1	1	2	1	—	—	—	1	—	—	—	—	1	
1913	10	5	2	1	3	1	2	1	2	2	—	2	2	2	—	1	1	—	1	—	—	1	—	—	—	—	—	—	1	—	
1914	9	6	3	4	1	2	1	1	1	—	2	2	—	1	1	—	—	1	—	—	—	—	—	1	—	2	—	—	—	—	
1915	16	5	4	1	1	—	1	—	—	—	—	—	—	2	—	1	—	—	1	—	1	—	—	—	1	—	—	—	1	—	
1916	15	8	2	1	1	1	—	—	1	1	1	1	1	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
1917	11	6	5	6	5	3	1	2	1	1	—	—	—	—	1	—	—	—	1	1	—	2	—	—	—	—	—	—	—	—	
1918	13	5	4	1	1	—	2	—	1	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
1919	12	2	2	3	1	—	2	1	2	2	—	—	1	1	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	
1920	15	5	—	2	1	—	1	1	1	—	—	1	—	—	—	—	—	1	—	—	1	—	—	—	—	—	—	—	—	—	
1921	15	3	3	2	1	2	—	—	—	—	1	—	2	—	—	—	1	—	—	—	—	—	—	—	—	—	1	1	—	—	
1922	6	5	3	1	2	3	—	1	—	1	1	1	—	—	—	—	—	—	—	4	—	—	1	—	—	—	—	—	—	
1923	6	1	1	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
1924	11	3	2	—	1	2	1	—	5	1	1	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	
1925	18	4	—	—	2	—	1	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	
1926	13	4	2	2	—	1	1	1	—	—	1	—	—	1	1	—	—	1	—	1	2	—	—	—	—	—	—	—	—	
1927	14	5	3	2	3	2	1	—	—	1	1	1	—	—	—	—	—	—	2	—	—	—	—	—	1	—	2	—	—	
1928	12	2	2	1	—	1	—	1	1	—	—	1	—	—	—	—	—	1	—	—	—	—	—	—	—	1	—	—	—	
1929	10	4	1	1	2	2	2	1	2	—	1	1	2	—	—	1	—	1	—	—	—	—	—	1	—	—	—	—	—	
1930	22	6	3	4	7	2	—	—	2	1	—	1	1	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	
1931	15	7	7	4	—	—	2	1	2	1	1	3	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	
1932	11	6	1	2	—	2	—	—	—	—	—	—	—	—	2	2	1	—	—	—	—	—	—	—	—	—	—	—	—	
1933	9	7	1	3	1	—	2	—	—	—	—	—	—	1	—	—	—	2	—	2	—	—	—	—	—	—	—	—	—	
1934	11	4	3	2	—	1	1	—	—	1	—	3	—	—	1	1	1	—	—	—	—	—	—	—	—	—	—	—	1	—	
1935	5	2	2	1	—	1	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	
1936	9	—	—	1	1	2	—	3	1	—	1	—	1	—	—	—	1	—	—	—	—	—	—	1	—	—	—	—	—	
	363	132	80	62	45	32	28	22	27	16	13	19	12	11	9	10	7	9	8	10	7	4	2	4	3	7	4	2	5	3	

Tableau XIX

mmagées dans chaque jour de grêle (Bade, 1908-36)

IUNES ENDOMMAGÉES A ÉTÉ LE SUIVANT :

37	38	40	41	43	45	46	47	48	49	50	51	52	54	56	57	58	60	61	62	63	68	74	79	81	92	96	116	Nombre total
—	—	—	—	—	1	1	—	—	1	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	1	40
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	—	38
—	—	—	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	58
—	—	—	—	—	1	—	—	1	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	38
—	1	—	1	—	—	—	—	2	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	54
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	40
1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	42
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	35
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	36
—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	48
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	28
—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	33
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	29
—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	35
—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	32
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	10
—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	29
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1	—	—	—	—	—	—	—	30
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	33
—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	39
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	25
—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	35
—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	1	—	1	—	—	—	54
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	45
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	28
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	28
—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	31
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	13
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	21
3	3	2	1	2	3	2	2	3	1	—	2	1	—	1	2	1	1	—	1	2	2	2	1	1	1	1	1	1.007

dommages de tous les cas d'incendie d'après les statistiques du Canton de Berne.

Pour les dits cas d'incendie grave, l'indice de dispersion des montants des dommages dans les dix périodes décennales du siècle 1806-1905 est $Q_7 = 1{,}94$. Cette valeur est comprise entre celle ($Q_2 = 3{,}14$) obtenue pour le nombre des cas d'incendie grave et celle ($Q''_A = 0{,}73$) que l'on obtient pour les moyennes des dommages pour chaque cas d'incendie grave.

De même, pour tous les cas d'incendie considérés par les statistiques du Canton de Berne pendant les cent années 1807-1906, les valeurs de Q_7 (1807-34 : 1,25 ; 1835-58 : 1,39 ; 1859-82 : 1,67 ; 1883-1906 : 1,03) sont inférieures aux valeurs respectives de Q_2 obtenues pour les cas d'incendie (1,36 ; 3,06 ; 3,91 ; 2,39) ; mais supérieures à celles de Q''_A que l'on obtient pour les dommages moyens pour chaque cas d'incendie (respectivement : 1,05 ; 1,06 ; 1,05 ; 0,98).

*III*ème *application : Grêle.*

Dans la thèse de doctorat de M. W. Rohrbeck présentée en 1909 à l'Université de Berlin, que nous avons déjà utilisée (cf. page 37), on trouve, à la page 15, non seulement les données sur le nombre des jours de grêle dans le pays de Bade, auxquelles nous avons appliqué le schéma de la dispersion du premier ordre ; mais aussi des données sur le nombre des communes endommagées, ainsi que sur la superficie endommagée dans les mêmes années 1868-1907.

Des données semblables pour les années 1908-1936 m'ont été communiquées, par lettre du 5 novembre 1941, par le Directeur du Bureau de Statistique de Bade, qui ensuite (lettres des 8 et 24 octobre 1942) m'a aussi communiqué les distributions : *a*) du nombre des communes endommagées, et *b*) des superficies endommagées dans chaque jour de grêle pour les mêmes années 1908-1936. Ces données, qui sont exposées aux tableaux XVIII et XIX, permettent l'application du schéma de dispersion du deuxième ordre.

On obtient la valeur de $Q_5 = 1{,}57$, pour la série des nombres des communes endommagées, et de $Q_5 = 1{,}23$ pour la

série des superficies endommagées (1). C'est-à-dire que la dispersion se trouve être légèrement au-dessus de la normale.

L'indice de dispersion Q''_A, pour les nombres moyens des communes frappées dans un jour de grêle, est = 1,11 et, pour la superficie moyenne frappée dans un jour de grêle = 0,98.

Rappelons que l'application de la formule (3.1) avait porté, pour les jours de grêle, à un indice de dispersion $Q_3 = 1,95$.

Les valeurs de Q_5 sont donc intermédiaires entre celles de Q'_3 et de Q''_A, aussi bien pour le nombre des communes que pour les superficies endommagées.

IV^ème application : Maladies des employés publics assurés.

Nous avons fait (cf. pages 40-41) une application du schéma de la dispersion du premier ordre aux cas de maladie qui se sont produits dans les différentes régions d'Italie parmi les inscrits à l'ENPDEP et, aux indemnités correspondantes qui ont été liquidées en 1952, en obtenant une valeur de $Q_2 = 5,0$.

Pour les mêmes cas de maladie, le service statistique de cet institut m'a aimablement communiqué la distribution des cas de maladie d'après leur durée, ainsi que des indemnités liquidées d'après leur montant pour les diverses régions. Les données sont exposées au tableau XX. Elles permettent l'application du schéma de dispersion du deuxième ordre.

On obtient les valeurs suivantes de Q_6 :

pour la durée des maladies, 6,9

pour les indemnités liquidées, 2,2.

Les valeurs respectives de Q''_A sont 11,4 et 2,7, tandis que la valeur de l'indice de dispersion du premier ordre pour la fréquence des maladies est, ainsi que nous l'avons vu, $Q_2 = 5,0$.

La valeur de Q_6, pour ce qui concerne les indemnités ($Q_6 = 6,9$), est donc intermédiaire entre celles de $Q''_A = 11,4$ et de $Q_2 = 5,0$, bien qu'étant beaucoup plus proche de la valeur inférieure, tandis que, pour les durées des maladies, la valeur de Q_6 ($= 2,2$) reste au-dessous des deux valeurs de Q_2 ($= 5,0$)

(1) Dans ces calculs, la même commune, ou la même superficie, a été comptée plusieurs fois, si dans l'année elle a été frappée plusieurs fois par la grêle.

Nombres et durée des cas de maladie et indemn

RÉGIONS	NOMBRES DES INSCRITS	NOMBRES DES CAS DES MALADIE	NOMBRE DES CAS DE MALADIE QUI ONT EU LA DURÉE INDIQUÉE CI-DESSOUS (EN JOURS)						
			Jusqu'à 3	4-7	8-15	16-30	31-60	61-90	au-de de
Piémont (y compris le val d'Aoste)	8.197	758	92	106	250	144	113	34	
Lombardie	11.818	989	90	132	323	221	160	40	
Trentin, Haut-Adige	1.892	207	11	12	73	52	41	11	
Vénétie.	8.152	798	34	52	247	226	164	46	
Frioul, Vénétie Julienne. . .	1.603	169	10	14	54	35	39	10	
Ligurie.	5.211	490	27	40	116	180	83	27	
Emilie et Romagne	10.495	1.132	52	71	354	309	241	65	
Toscane.	9.880	1.012	74	70	294	236	224	74	
Ombrie.	1.724	141	5	8	44	39	32	7	
Marches.	2.977	255	12	10	54	67	78	24	
Latium	34.018	3.417	1.005	381	923	602	373	93	
Abruzzes et Molise.	2.570	185	12	9	48	49	41	12	
Campanie.	9.599	1.172	173	191	354	182	188	63	
Pouille	6.607	564	14	20	165	142	157	35	
Basilicate.	987	58	2	5	19	17	11	4	
Calabrie.	2.940	197	7	9	51	47	49	24	
Sicile.	9.312	576	41	49	145	126	131	57	
Sardaigne.	2.235	97	6	9	27	17	30	6	
ITALIE . . .	130.217	12.217	1.667	1.188	3.541	2.691	2.155	632	

Pour calculer la durée totale des maladies des diverses classes, on a admis que la durée moyenne d'une maladie était: de à la demi-somme des limites de la classe, pour toutes les autres classes (par ex., de 45 jours pour les maladies de la l'indemnité moyenne était : de 1500 lires pour la classe des indemnité jusqu'à 1.999 lires ; de 96.000 lires pour la de autres classes (p. ex., de 17.500 lires pour la classe des indemnite de 15.000 a 19.999 lires). Les moyennes de la première

Tableau XX

...uidées pendant l'année 1952 *à des inscrits à l'ENPDEP*

...squ'à 999	2.000-3.999	4.000-5.999	6.000-7.999	8.000-9.999	10.000-14.999	15.000-19.999	20.000-29.999	30.000-49.999	50.000 et au-dessus
	Nombres des cas de maladies qui ont donné lieu à une indemnité se montant au chiffre indiqué ci-dessous (en lires)								
199	195	92	81	50	51	32	20	18	20
165	226	155	92	81	99	40	41	40	50
26	53	35	24	12	26	9	8	12	2
96	178	166	82	58	96	35	31	34	22
15	27	27	19	16	15	11	9	21	9
74	104	76	60	36	56	31	22	12	19
152	234	187	131	97	147	58	44	41	41
189	230	186	94	67	101	55	41	26	23
12	30	18	19	9	21	13	12	5	2
26	37	33	32	26	46	24	15	7	9
559	780	553	347	246	342	166	142	154	128
22	27	35	22	15	24	10	11	12	7
230	252	171	124	84	129	68	60	43	11
20	63	80	83	56	107	62	51	24	18
4	8	11	8	6	4	8	6	2	1
13	20	29	24	22	27	18	19	19	6
46	70	67	73	42	119	56	53	32	18
13	10	14	12	8	19	3	10	5	3
.861	2.544	1.935	1.327	931	1.429	699	595	507	389

..., pour la classe des maladies jusqu'à 3 jours ; de 113,5 jours, pour la classe des maladies au-dessus de 90 jours ; et égale ...1-60 jours). Analoguement, pour calculer le montant total des indemnités pour les diverses classes, on a admis que ...e des indemnités de 50.000 lires et au-dessus, et égale à la demi-somme des limites de la classe pour toutes les ...e la dernière classe ont été calculées sur la base de renseignements tirés d'autres publications de l'ENPDEP.

et de Q''_A ($= 2,7$). Ce dernier résultat s'explique par la corrélation négative qui existe entre fréquence des cas de maladie et montant des indemnités ; les régions dans lesquelles on est plus large en ce qui concerne les déclarations et l'admission des cas de maladie étant celles dans lesquelles les maladies sont plus légères, et par conséquent plus courtes, et comportent des indemnités moins importantes.

La corrélation négative est même plus élevée ($r = -0,82$) pour les durées que pour les indemnités ($r = -0,12$), car pour celles-ci il y a un facteur de perturbation dans leur corrélation avec la fréquence des cas de maladie, provenant du pourcentage des cas hospitalisés, qui, à égalité de durée de la maladie, sont bien plus coûteux.

22. Dans les paragraphes précédents nous avons parlé de dispersion d'un ordre supérieur, mais, en réalité, nous n'avons fourni que des exemples de dispersion du deuxième ordre. Il est aisé d'apporter des exemples de grandeurs énumérables ou mesurables dont on peut étudier la dispersion de l'ordre troisième, quatrième, etc.

La série des nombres des tremblements de terre peut être traitée suivant le schéma de la dispersion du premier ordre, et la série des nombres des secousses sismiques, ainsi que nous l'avons dit, suivant le schéma de la dispersion du deuxième ordre, dans lequel la deuxième urne contient des boules avec la mention du nombre moyen des secousses pour chaque tremblement. La série des nombres des bâtiments endommagés par les tremblements de terre peut être traitée suivant le schéma de la dispersion du troisième ordre, dans lequel la troisième urne contient les mentions relatives au nombre moyen des bâtiments endommagés par chaque secousse. La série des montants des dommages causés par les tremblements de terre, enfin, peut être traitée suivant le schéma de la dispersion du quatrème ordre, la quatrième urne fournissant les indications relatives au dommage moyen causé à chacun des bâtiments endommagés.

De façon analogue, la série du nombre des assurés qui sont tombés malades pendant une année peut être traitée, aux fins de la dispersion, par le schéma de la dispersion du premier or-

dre ; la série de leurs maladies par le schéma de la dispersion du deuxième ordre, la deuxième urne fournissant les indications relatives au nombre moyen des maladies par malade ; la série des journées d'absence pour cause de maladie par le schéma de la dispersion du troisième ordre, la troisième urne fournissant les indications relatives au nombre moyen des jours d'absence pour chaque maladie ; la série des dommages qui en sont dérivés à l'administration par le schéma de la dispersion du quatrième ordre, la quatrième urne fournissant les indications relatives au dommage moyen que l'administration a subi pour chaque jour d'absence.

23. Il est évident qu'une même série peut être traitée par des schémas d'ordres divers, puisque la grandeur résultante peut être scindée en un nombre divers de composantes. Nous avons vu que le montant des dommages pour cause de maladie peut être scindé en quatre composantes, mais il pourrait être scindé aussi en trois composantes, auxquelles, aux fins de la mesure de la dispersion théorique, correspondraient trois urnes : nombre des cas de maladie, nombre moyen des maladies par malade, moyenne des dommages que l'administration subit pour chaque maladie, ou bien aussi : nombre des malades, nombre moyen des absences par malade, moyenne des dommages pour chaque jour d'absence ; ou encore : nombre des cas de maladie, nombre moyen des jours d'absence pour maladie, moyenne des dommages pour chaque jour de maladie. Et il pourrait aussi être scindé en deux composantes, auxquelles correspondraient deux urnes qui fourniraient les données concernant le nombre des malades et la moyenne des dommages pour chaque malade, ou bien concernant le nombre des absences et la moyenne des dommages pour chaque absence, ou encore concernant le nombre des cas de maladie et la moyenne des dommages pour chaque maladie.

24. Remarquons que les différentes composantes qui interviennent dans l'étude de la dispersion d'un ordre supérieur peuvent être toutes limitées ou bien toutes illimitées ou, enfin, en partie illimitées et en partie limitées. Par exemple, est illimité le nombre des précipitations atmosphériques ainsi que la quantité de l'eau tombée ; est limité, par contre, le nombre des jours

où des précipitations se produisent, mais est illimitée la quantité de l'eau tombée par jour. Sont limités, au contraire, le nombre des jours de grêle ainsi que les dommages causés par elle aux récoltes. Illimités sont soit le nombre des tremblements de terre et soit le nombre des secousses pour chaque tremblement ; sont limités, par contre, le nombre des bâtiments endommagés et le montant des dommages qui y ont été causés. Est limité le nombre des malades, qui ne peut dépasser le nombre des individus exposés à tomber malades, et est illimité le nombre de cas de maladie, mais est limité le nombre des jours d'absence ainsi que le montant des dommages que l'administration peut subir par suite de cas d'absence.

25. Avant de clore cet article, je crois utile de faire quelques observations au sujet des différentes formules que nous avons données pour l'indice de dispersion des moyennes.

J'entends parler des formules (9), (10), (11) (dont au paragraphe 18) qui donnent les valeurs des indices de dispersion des moyennes et que nous avons indiqués par les notations Q_A, Q'_A, Q''_A. Ces formules diffèrent entre elles uniquement par les diverses pondérations que l'on adopte pour les carrés des écarts $(A_k - A)$.

Pour la détermination de Q_A, on attribue à tous ces écarts le même poids ; pour la détermination de Q'_A, on leur attribue un poids proportionnel aux nombres m_k des observations dont la moyenne A_k est tirée, et, enfin, pour la détermination de Q''_A, un poids proportionnel au carré m_k^2 du dit nombre des observations.

Comme la valeur probable du carré de l'écart diminue proportionnellement au nombre des observations, il paraît justifié de multiplier ce carré par le dit nombre d'observations afin de donner à tous les termes la même importance. C'est ce qu'on fait dans la formule Q'_A ; mais on ne voit pas une raison théorique pour laquelle on devrait s'attendre à ce que les valeurs de Q_A, Q'_A, Q''_A diffèrent systématiquement.

Il est donc intéressant de voir comment les choses se passent dans la réalité.

Dans le tableau XXI, sont classées les valeurs de Q_A, Q'_A Q''_A pour les onze séries de données qui en ont permis le calcul.

Dans sept séries de données (col. 3, 6, 7, 8, 9, 10, 12), les valeurs des trois indices diffèrent très peu entre elles ; dans trois autres, (col. 2, 4, 5) elles diffèrent sensiblement ; dans une, enfin (col. 11), elles diffèrent essentiellement.

Dans les sept séries pour lesquelles les indices diffèrent très peu, les valeurs augmentent en deux séries (col. 9, 12) en passant de Q'_A à Q''_A ; elles diminuent au contraire en trois autres séries (col. 6, 10, 7) et ne présentent pas une tendance nette dans les deux autres (col. 3, 8).

Dans les quatre séries pour lesquelles les indices diffèrent plus ou moins sensiblement, leur valeur augmente en passant de Q_A à Q''_A.

L'augmentation la plus marquée se produit pour la série de la durée moyenne des cas de maladie dans les diverses régions de l'Italie (col. 11). Elle est due essentiellement à la situation tout à fait spéciale du Latium, où le nombre des cas de maladie est très élevé et leur durée moyenne est très basse. Il est probable que, dans les trois autres séries aussi, des anomalies analogues, bien que moins sensibles, expliquent les divergences entre les diverses valeurs de l'indice de dispersion des moyennes.

Aux deux dernières lignes du tableau XXI on trouvera les valeurs :

a) des coefficients de corrélation ($r_{A_k m_k}$) entre les moyennes (A_k) dont on a mesuré la dispersion et les nombres (m_k) des cas sur lesquels les dites moyennes sont basées

b) des coefficients de corrélation ($r_{e_k m_k}$) entre les écarts (e_k) des dites moyennes A_k de la moyenne générale et les dits nombres des cas.

Dans trois séries (col. 6, 7 et 10), dans lesquelles la valeur de l'indice de dispersion diminue en passant de Q'_A à Q''_A, la valeur de $r_{e_k m_k}$ est négative ; dans les six séries dans lesquelles, au contraire, la valeur du dit indice de dispersion augmente, quatre fois (col. 2, 4, 5 et 11) la valeur de r est positive et deux fois (col. 9 et 12) négative.

Il paraît donc que, dans la plupart des cas (sept fois sur neuf), il y a accord entre le signe positif ou négatif du coefficient de corrélation $r_{ek m_k}$ et l'augmentation, ou respectivement la diminution, de la valeur de l'indice de dispersion en passant de Q'_A à Q''_A.

Tableau XXI

Valeurs des indices de dispersion des moyennes et coefficients de corrélation

INDICES DE DISPERSION	Secousses des tremblements de terre	Incendies						Grêle		Cas de maladie	
		Nombre des bâtiments endommagés (Argovie)	Montant des dommages (Argovie)	Montant des dommages (Berne)				Nombre des communes endommagées	Superficie endommagée	Durée de la maladie	Indemnités liquidées
				1807-34	1935-38	1859-82	1883-1906				
1	2	3	4	5	6	7	8	9	10	11	12
Q'_A	1,93	2,72	0,41	0,61	1,22	1,26	0,987	1,08	1,11	3,91	2,04
Q''_A	2,17	2,64	0,57	0,89	1,16	1,17	0,990	1,08	1,02	6,51	2,64
Q'''_A	2,75	2,75	0,73	1,05	1,06	1,05	0,982	1,11	0,98	11,42	2,72
Coefficients de correlation											
$r_{A_k m_k}$	+0,71	−0 76	+0,44	+0,24	+0,01	+0,28	−0 26	+0,33	−0,08	−0,82	−0,12
$r_{e_{A_k} m_k}$	+0,48	−0 50	+0,08	+0,19	−0,19	−0,54	−0 19	−0,13	−0,40	+0,55	−0,81

Au contraire, on ne remarque aucune relation entre le signe positif ou négatif de l'indice de corrélation $r_{A_k m_k}$ et la variation de l'indice de dispersion en passant de Q'_A à Q''_A.

25. Les considérations développées dans cet article montrent que le concept de dispersion peut être convenablement étendu de façon à devenir applicable à des séries de grandeurs auxquelles on n'avait pas pensé par le passé. Elles montrent pareillement que la dispersion d'une série de grandeurs énumérables et mesurables peut souvent être mesurée de diverses façons, de sorte que les concepts de dispersion normale ou supernormale ou sous-normale, n'acquièrent de signification qu'en relation avec le schéma sur la base duquel la dispersion est mesurée. Des questions collatérales ont été aussi examinées et des applications multiples des formules proposées ont été faites, qui ont porté à des résultats intéressants à plusieurs points de vue.

Résumé

L'auteur expose d'abord les critériums sur lesquels est fondé l'indice de *dispersion d'après le schéma des rapports de composition*. A la différence de l'indice de dispersion usuel, fondé sur les rapports de dérivation et qui représente la mesure de la dispersion applicable à des fréquences relatives (par ex., taux de mortalité, rapports de masculinité, pourcentage des édifices endommagés) ou à des fréquences absolues illimitées (par ex., nombre des décès, des nouveau-nés du sexe masculin, des édifices incendiés), cet indice nouveau représente la mesure de la dispersion applicable à des fréquences absolues limitées, et il permet donc d'étendre la théorie de la dispersion aux grandeurs absolues énumérables (par ex., nombre des tremblements de terre, des poissons pêchés). Les applications à des phénomènes divers mettent en lumière de remarquables régularités statistiques et font voir que la dispersion augmente fortement lorsque l'on passe des phénomènes météorologiques aux phénomènes sur lesquels influe l'intervention de l'homme et qu'elle atteint les valeurs les plus élevées à l'égard des phénomènes qui se ressentent de facteurs physiques aussi bien que de la conduite de l'homme et d'autres espèces animales. L'indice de dispersion apparaît, en outre, plus élevé pour les séries comprenant un plus grand nombre d'années, et ce par l'effet de tendances systématiques qui dans un moindre nombre d'années sont moins évidentes et, dans le cas de courts intervalles, peuvent même être neutralisées par des tendances compensatrices.

L'auteur illustre, ensuite, l'extension ultérieure de la théorie de la dispersion au moyen du schéma de la *dispersion d'un ordre supérieur*, dans lequel le phénomène dont on étudie la dispersion est assimilé à un schéma d'extractions successives, avec ou sans remise, de deux urnes, ou plus, qui correspondent aux diverses composantes du phénomène. Ce schéma s'applique soit aux grandeurs mesurables à composantes énumérables, que l'auteur nomme *montants extensifs* (par ex., montant des dommages causés par les incendies, par la grêle, etc.), soit aux grandeurs énumérables à composantes énumérables, ou *montants intensifs* (par ex., nombre des secousses de tremblement de terre). Ces derniers

montants peuvent être traités aussi suivant le schéma de la dispersion du premier ordre. Il est ainsi démontré, sur le fondement d'exemples concrets, que les concepts de dispersion normale, sous-normale et supernormale acquièrent un sens précis seulement alors que le schéma sur le fondement duquel on juge de la dispersion est précisé.

Summary

The author first sets forth the criteria on which the *dispersion* index is founded *according to the scheme of the composition ratios*. Unlike the usual dispersion index, founded on derivation ratios and which gives the measure of dispersion applicable to relative frequencies (for example, mortality rates, masculinity ratios, percentage of buildings damaged), or to limited absolute frequencies (for example, number of deaths, of newborn males, of buildings catching fire), this new index gives the measure of the dispersion applicable to unlimited absolute frequencies and thus allows the extension of the dispersion theory to absolute enumerative quantities (for example, number of earth-quakes, of fish caught). The applications made to several phenomena bring to light some remarkable statistical regularities and show that dispersion increases greatly when from meteorological phenomena we pass on to phenomena upon which the interference of man has some influence, and it reaches the highest values in the case of phenomena that are affected by physical factors as well as by the action of man and other animal species. The dispersion index, besides, is higher in the case of series including a large number of years, owing to the effect of systematic tendencies which in a lesser number of years are less evident and in short intervals may even be neutralized by compensatory tendencies.

The author illustrates, then, the further extension of the dispersion theory by means of the scheme of the *dispersion of a higher order*, in which the phenomenon whose dispersion is being investigated is assimilated to a scheme of successive extractions, with or without replacement, from two or more boxes corresponding to the different components of the phenomenon. This scheme

applies both to measurable quantities with enumerative components, called by the author *extensive amounts* (for example, amount of damages caused by fires, by hail-storms) and to enumerative quantities with enumerative components, or *intensive amounts* (for example, number of earthquake shocks). The latter amounts may be treated also according to the scheme of the dispersion of the first order. It is thus demonstrated, on the basis of concrete examples, that the concepts of normal, sub-normal and supernormal dispersion acquire a precise meaning only if the scheme on whose ground one judges of the dispersion is specified.

Zusammenfassung

Verfasser erklärt zuerst die Kriterien auf welchen der *Dispersions-index mit Rücksicht auf das « Schema der Komponentenverhältnisse* begründet ist. Zum Unterschiede vom gewöhnlichen Dispersionsindex, der auf den Derivationsverhältnissen begründet ist und das Mass der Dispersion angibt das anzuwenden ist auf die relativen Frequenzen (z. B. Sterblichkeit, Verhältnis der männlichen Geburten, Prozentsatz der beschädigten Gebäude) oder auf die begrenzten absoluten Frequenzen (z. B. Zahl der Todesfälle, der männlichen Geburten, der in Brand geratenen Gebäude), dieser neuartige Index ergibt das Mass der Dispersion bei das auf die unbegrenzten absoluten Frequenzen anzuwenden ist, und ermöglicht also die Ausdehnung der Dispersionstheorie auf die aufzählbaren absoluten Grössen (z. B. Zahl der Erdbeben, der gefangenen Fische). Die Anwendungen auf verschiedenartige Erscheinungen bringen ans Licht bemerkenswerte statistische Regelmässigkeiten und zeigen, dass die Dispersion stark zunimmt, wenn man von den meteorologischen Phänomenen zu Phänomenen übergeht, welche durch den Menschen beeinflusst werden, und die höchsten Werte erreicht im Falle von Phänomenen, die dem Einfluss sowohl von physischen Faktoren als auch vom Betragen des Menschen und anderer animalischer Arten unterliegen. Ferner erscheint der Dispersionsindex böher bei den eine grössere Anzahl von Jahren umfassenden Serien, und zwar

durch die Wirkung von systematischen Tendenzen, die in einer geringeren Anzahl von Jahren wenig angenscheilich sind und in kürzeren Zwischenräumen sogar von ausgleichenden Tendenzen neutralisiert werden können.

Verfasser erläutert dann die weitere Ausdehnung der Theorie durch das Schema der *Dispersion höherer Ordnung*, bei dem das Phänomen, dessen Dispersion man untersucht, assimiliert wird einem Schema von aufeinander folgenden Ziehungen, mit oder ohne Wiederholung, aus zwei oder mehr Gefässen, die den verschiedenen Komponenten des betreffenden Phänomens entsprechen. Dieses Schema wird angewendet sowohl auf die messbaren Grössen mit aufzählbaren Komponenten, die der Verfasser *extensiven Beträge* nennt (z. B. Betrag der durch Feuer, Hagelwetter u.s.w. verursachten Schäden), wie auch auf die aufzählbaren Grössen mit aufzählbaren Komponenten, nämlich die *intensiven Beträge* Cz. B. Zahl der Erdbebenstösse). Letztere Beträge können auch behandelt werden auf Grund des Dispersionsindex erster Ordnung. So wird mit konkreten Beispielen bewiesen, dass die Begriffe von normaler, unternormaler und übernormaler Dispersion eine genaue Bedeutung nur dann gewinnen, wenn das Schema genau angegeben ist, auf dessen Grundlage die Dispersion beurteilt wird.

Riassunto

L'A. espone anzitutto i criteri su cui si fonda l'indice di *dispersione in base allo schema dei rapporti di composizione*. A differenza dell'usuale indice di dispersione, fondato sui rapporti di derivazione, che rappresenta la misura della dispersione da applicarsi alle frequenze relative (per es., tassi di mortalità, rapporti di mascolinità, percentuale degli edifici danneggiati) o di frequenze assolute limitate (per es. numero dei morti, dei nati maschi, degli edifici incendiati), questo nuovo indice rappresenta la misura della dispersione da applicarsi alle frequenze assolute illimitate e permette quindi di estendere la teoria della dispersione alle grandezze enumerabili assolute (per es., numero dei terremoti, dei pesci pescati). Le applicazioni a svariati fenomeni mettono in luce note-

voli regolarità statistiche e mostrano che la dispersione aumenta fortemente quando si passa dai fenomeni meteorologici a fenomeni sui quali influisce l'intervento dell'uomo e raggiunge i valori più alti per fenomeni che risentono sia di fattori fisici che della condotta dell'uomo e di altre specie animali. L'indice di dispersione risulta inoltre più elevato per le serie comprendenti un maggior numero di anni, per effetto di tendenze sistematiche che in un numero minore di anni risultano meno manifeste e in intervalli brevi possono anche essere neutralizzate da tendenze compensatrici.

L'A. illustra poi l'ulteriore estensione della teoria della dispersione mediante lo schema della *dispersione di ordine superiore*, nel quale il fenomeno di cui si studia la dispersione viene assimilato ad uno schema di estrazioni successive, con o senza ripetizione, da due o più urne che corrispondono alle diverse componenti del fenomeno. Questo schema si applica sia alle grandezze misurabili a componenti enumerabili, i cosiddetti *ammontari estensivi* (per es. ammontare dei danni prodotti dagli incendi, dalla grandine, ecc.), sia alle grandezze enumerabili a componenti enumerabili, o *ammontari intensivi* (per es., (1) numero delle scosse di terremoto). Questi ultimi possono essere trattati anche secondo lo schema della dispersione di primo ordine. Si dimostra così, sulla scorta di esempi concreti, che i concetti di dispersione normale, subnormale e supernormale acquistano un significato preciso solo se è precisato lo schema probabilistico in base al quale si giudica della dispersione.

---

(1) In italiano è stata usata, in luogo dell'espressione *ammontare intensivo*, la parola *numerario*, riserbando la parola *ammontare* a designare, secondo il significato corrente, l'*ammontare estensivo* (cfr. l'articolo citato *Estensioni e portata, ecc.*, « Giornale dell'Istituto Italiano degli Attuari », 1955, pag. 11. Non si è saputo trovare una traduzione esatta della parola *numerario* nelle altre lingue, cosicchè in questo articolo si è preferito parlare sempre di *ammontare intensivo* e, in contrapposto, di *ammontare estensivo*.

F. N. DAVID and N. L. JOHNSON

Reciprocal Bernoulli and Poisson variables

1. In the course of research on censored distributions we have encountered the reciprocal Bernoulli variable, defined by the distribution function

$$P\{x = r^{-1}\} = \binom{n}{r} p^r q^{n-r} (1 - q^n)^{-1} \quad (r = 1, 2, \ldots n) \qquad (1)$$

where

$$0 < p < 1 \quad \text{and} \quad p + q = 1.$$

Taking the work of Stephan (1945) as a starting point we have obtained approximate formulae for moments of x which are of interest. In the present note these formula will be compared with results given by Oliveira (1952) and with numerical checks provided by Fieller and Hartley (1954). More extensive tables are provided by Grab and Savage (1954).

2. The s^{th} moment of x about zero is $\mu'_s = E(x^s) = E(y^{-s})$ where $y = x^{-1}$. The moments of y, $\widetilde{m}'_s$, are simply the corresponding Bernoulli moments m'_s divided by $(1 - q^n)$. We also note the formula

$$E[(y - n p)^s] = [m_s + (-1)^{s-1} (n p)^s q^n] (1 - q^n)^{-1} \qquad (2)$$

where m_s denotes a central Bernoulli moment.

$$\begin{aligned} E(x^s) &= E[\{n p + (y - n p)\}^{-s}] \\ &= (n p)^{-s} E\left[\left(1 + \frac{y - n p}{n p}\right)^{-s}\right] \end{aligned} \qquad (3)$$

We now expand the quantity in the expectation and take expected values of each term in the summation. The result so obtained may be written formally as follows

$$E(x^s) = (np)^{-s} \sum_{r=0}^{\infty} (-1)^r \binom{s+r-1}{r} \frac{m_r + (-1)^{r-1} (np)^r q^n}{(np)^r (1-q^n)} \qquad (4)$$

Curtailing this series at the $(R+1)$ th term we obtain the approximate formula

$$E(x^s) \doteqdot (np)^{-s} (1-q)^{-1} \cdot \left\{ \sum_{r=0}^{R} (-1)^r \binom{s+r-1}{r} (np)^{-r} m_r - \binom{s+R}{R} q^n \right\}. \qquad (5)$$

If q^n is negligible we have the simpler approximate formula

$$E(x^s) \doteqdot (np)^{-s} \sum_{r=0}^{R} (-1)^r \binom{s+r-1}{r} (np)^r m_r. \qquad (6)$$

From (6) we obtain

$$\mu_1' = E(x) \doteqdot (np)^{-1} [1 + (np)^{-1} q + (np)^{-2} q(q+1) + + (np)^{-3} q(q^2 + 4q + 1) + (np)^{-4} q(q^3 + 11q^2 + 11q + 1] \qquad (7)$$

$$\mu_2 = var(x) \doteqdot (np)^{-3} q[1 + 2(np)^{-1}(2q + 1)]. \qquad (8)$$

These are not to be regarded as initial terms in expansions for μ_1', and μ_2, but as approximation formulae. Comparison with the tables of Fieller and Hartley shows that the first four terms of (7) give values correct to four significant figures for $p > \frac{2}{5}$ and $n > 25$; the first three terms suffice for $p > \frac{3}{5}$ and $n > 25$; and the first two for $p > \frac{4}{5}$, $n > 50$. Formula (7) is poor when q^n is not negligible. In such cases (5) is preferable.

To obtain corresponding formulae for a reciprocal Poisson variable with mean ξ we may put $np = \xi$ above and let $n \to \infty$,

$p \to 0$, $q \to 1$. Alternatively we may work directly from the distribution. In either case we find, analogously to (5)

$$\mu_s' \doteqdot \xi^{-s}(1-e^{-\xi})\left[\sum_{r=0}^{R}(-1)^r\binom{s+r-1}{r}\xi^{-r}m_r - \binom{s+R}{R}e^{-\xi}\right] \qquad (9)$$

If $e^{-\xi}$ is negligible, then analogously to (7) and (8),

$$\mu_1' \doteqdot \xi^{-1}(1+\xi^{-1}+2\xi^{-2}+6\xi^{-3}+24\xi^{-4}) \qquad (10)$$

$$\mu_2 \doteqdot \xi^{-3}(1+6\xi^{-1}) \qquad (11)$$

Formula (10) gives results correct to four significant figures if $\xi > 25$; the first four terms give results correct to three figures if $\xi > 10$. The formula is not useful for small values of ξ, and (9) is then to be preferred.

3. Oliveira (1952) has derived the differential equation

$$\frac{d\,\mu_1'}{d\,q} = \frac{n}{q(1-q^n)}\mu_1' - \frac{1}{q(1-q)} \qquad (12)$$

for the mean of the reciprocal Bernoulli variable. Solving (12) we obtain

$$(q^{-n}-1)\,\mu_1' = A - \int q^{-(n+1)}(1-q)^{-1}(1-q^n)\,d\,q$$

where A is a constant chosen to make $\mu_1' \to 1$ as $q \to 1$.

Hence

$$\mu_1' = \frac{1}{1-q^n}\sum_{r=0}^{n-1}\frac{q^r-q^n}{n-r} \qquad (13)$$

A similar formula can easily be reached by noting that

$$\sum_{r=1}^{n}\binom{n}{r}r^{-1}q^r p^{n-r} = \sum_{r=1}^{n-1}\binom{n-1}{r}r^{-1}p^r q^{n-r} + n^{-1}\sum_{r=1}^{n}\binom{n}{r}p^r q^{n-r}$$

whence, by repeated application, the result follows. Similarly we obtain the second moment about zero from the relation

$$\sum_{r=1}^{n}\binom{n}{r} r^{-2}\, p^r\, q^{n-r} = \sum_{r=1}^{n-1}\binom{n-1}{r} r^{-2}\, p^r\, q^{n-r} +$$

$$+ n^{-1}\sum_{r=1}^{n-1}\binom{n-1}{r} r^{-1}\, p^r\, q^{n-r} + n^{-2}\sum_{r=1}^{n}\binom{n}{r} p^r\, q^{n-r}$$

Unfortunately none of the quantities obtained in this way are easily summable since they are in effect the first n terms of various divergent logarithmic series. (13) will not be of much use for calculation purposes unless n is large and q is small.

The problem referred to in section 1 of this paper arises in connection with the type of censoring (termed Type I by Gupta (1952)) wherein all observations less than a certain fixed known value A are not recorded. If, in a sample of n individuals r are recorded then r will be a Bernoulli variable with

$$p = \int_A^\infty f(x)\, d\, x$$

where $f(x)$ is the population probability density function. If now, for example, we calculate the mean of the censored sample, it will have an expected value

$$p^{-1}\int_A^\infty x\, f(x)\, d\, x$$

and, *for a given* r, its variance will be

$$\pi^{-1}\left[p^{-1}\int_A^\infty x^2 f(x)\, d\, x - p^{-2}\left\{\int_A^\infty x\, f(x)\, d\, x\right\}^2\right] \qquad (14)$$

If $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ then (14) may be written

$$\pi^{-1}\left[1 + \frac{1}{p\sqrt{2\pi}} e^{-\frac{1}{2}A^2}\left\{A - \frac{1}{p\sqrt{2\pi}} e^{-\frac{1}{2}A^2}\right\}\right] = \pi^{-1} g(A), \text{ say.} \quad (15)$$

The variance, allowing for variation in r, of the mean of a sample censored in this way will be

$$g(A) \cdot E(r^{-1})$$

Some values of $n\, g(A) \cdot E(r^{-1})$ are shown in the following table.

A	p	$n\, g(A) \cdot E(r^{-1})$		
		$n = 10$	$n = 20$	$n = 50$
—2	0.977	0.909	0.908	0.908
—1.5	0.933	0.834	0.831	0.829
—1	0.841	0.765	0.756	0.752
—0.75	0.773	0.744	0.731	0.724
—0.5	0.691	0.744	0.721	0.710
—0.25	0.599	0.763	0.731	0.718
0	0.500	0.840	0.770	0.742

The existence of a minimum near $p = 0.60$ is noteworthy.

University College, London

REFERENCES

FIELLER, E. C. and HARTLEY, H. O. (1954), « Biometrika », 41, 494.
GRAB, E. L. and SAVAGE, I. R. (1954), « J. Amer. Statist. Assoc. », 49, 169.
GUPTA, A. K. (1952), « Biometrika », 39, 260.
OLIVEIRA, J. T. DE (1952), « An. Fac. Cien. Poto. », 36, 5.
STEPHAN, F. F. (1945). « An. Math. Statist. », 16, 50.

VITTORIO AMATO

On the distribution of Gini's *G* coefficient of rank correlation in rankings containing ties

Summary. — We want to examine the rank correlation problem in which ties are present in the rankings. For this tied case, we derive a generalised measure of rank correlation from which the Gini's coefficient follows as special case.

We show, moreover, a method of determining such a coefficient when the rankings are dichotomised.

The variance of Gini's coefficient, in a population of sample permutations, is given for the case where only one ranking contains some particular ties but the other does not.

Two practical applications of these results are given at the end of the paper.

1. — *Definition of Gini's coefficient of rank correlation.*

In this paper we consider the case when there are two rankings of n ranks and we want to examine the relationship between them.

Suppose, for instance, ten boys are ranked according to their merit in art and in science, as follows

Art :	3	5	1	2	4	6	7	9	8	10
Science :	1	3	5	4	6	2	9	7	10	8
Totals :	4	8	6	6	10	8	16	16	18	18

We need a synthetic measure of rank correlation among the two rankings based on dispersion of the sums 4, 8, 6, 6, 10, 8, 16, 16, 18, 18.

In the general case, we have

$$\xi_1 \quad \xi_2 \ldots . \xi_n$$

$$\eta_1 \quad \eta_2 \ldots . \eta_n$$

where ξ and η are two permutations of the first n integers.

We define a generalised coefficient of rank correlation by the equation (see : Amato, 1952-'53)

$$C = \frac{V(\xi + \eta) - V(\xi + \eta')}{V(2\eta)} . \qquad (1)$$

In this general expression we consider the quantity $V(\xi + \eta)$ as a generical measure of dispersion of the sum $\xi + \eta$. Similarly, we regard $V(\xi + \eta')$ as a measure of dispersion of the sum $\xi + \eta'$; and so for $V(2\eta)$. The rankings η and η' are "conjugate", such that $\eta + \eta' = n + 1$; and hence $V(\eta) = V(\eta')$.

In the case of *perfect concordance* among the rankings :

$$\xi = \eta , \quad V(\xi + \eta') = 0 , \quad C = +1 .$$

A necessary and sufficient condition for the *independence* of the rankings ξ and η, with respect to V, is that

$$V(\xi + \eta) = V(\xi) + V(\eta) .$$

It follows the result $C = 0$ in the case of independence.

Moreover, for *complete disagreement* between the same rankings :

$$\xi = \eta' , \quad V(\xi + \eta) = 0 , \quad C = -1 .$$

The coefficient C may vary therefore from -1 to $+1$, and corresponding to any positive value of C there is a negative value of the same magnitude arising from an inversion of one of the two rankings. The distribution of the values of C is symmetrical : a value of C obtained by correlating the ranking ξ with the other η and its conjugate η' are equal in magnitude but opposite in sign.

We may obtain, from (1), several coefficients of rank correlation based on mean deviation, on Gini's mean difference, on standard deviation, on variance and so on (see : Amato, 1952 - '53).

In particular, if we take the variance as a measure of dispersion of the sum $\xi + \eta$ and so for $\xi + \eta'$ and 2η, we find from (1) the well known rank correlation coefficient

$$\rho = \frac{var\,(\xi + \eta) - var\,(\xi + \eta')}{var\,(2\eta)} = \frac{cov\,(\xi, \eta)}{var\,\eta}$$

introduced by Spearman (1904).

If we assume the "sign-variance" (*) as a measure of dispersion we obtain from (1) the other well known rank correlation coefficient

$$\tau = \frac{\overline{var}\,(\xi + \eta) - \overline{var}\,(\xi + \eta')}{\overline{var}\,(2\eta)} = \frac{\overline{cov}\,(\xi, \eta)}{\overline{var}\,\eta}$$

due to Kendall (1938). In such a case, we write $\overline{var}$ instead of *var* and $\overline{cov}$ instead of *cov* to denote that we are considering

---

(*) We define a "sign-variance" for a permutation of the first n integers, the expression

$$\overline{var}\,\xi = \sum_{i>j} \alpha_{ij}^2 = \binom{n}{2}.$$

The sign-variance of the sum $\xi + \eta$ is given by

$$\overline{var}\,(\xi + \eta) = \overline{var}\,\xi + \overline{var}\,\eta + 2\,\overline{cov}\,(\xi, \eta),$$

where $\overline{cov}\,(\xi, \eta)$ is for "sign-covariance" between the rankings ξ and η, defined by

$$\overline{cov}\,(\xi, \eta) = \sum_{i>j} \alpha_{ij}\,\beta_{ij}$$

α_{ij}, β_{ij} being

$$\alpha_{ij} = +1 \qquad \xi_i > \xi_j$$
$$= -1 \qquad \xi_i < \xi_j$$

$$\beta_{ij} = +1 \qquad \eta_i > \eta_j$$
$$= -1 \qquad \eta_i < \eta_j$$

the scores -1 and $+1$ for any pair of ranks in the manner here indicated.

Finally, if we assume as a measure of dispersion the mean deviation about the mean, we get from (1)

$$G = \frac{\Sigma\,|\,\xi + \eta - (n+1)\,| - \Sigma\,|\,\xi + \eta' - (n+1)\,|}{\Sigma\,|\,2\,\eta - (n+1)\,|}\,, \qquad (2)$$

where the quantity $n + 1$ represents the arithmetic mean of the sum $\xi + \eta$, and so for $\xi + \eta'$ and $2\,\eta$. We see that the same value $n + 1$, is equal to $\eta + \eta'$; it follows, from substitution in (2), that G is equal to the well known coefficient of rank correlation

$$G = \frac{\Sigma\,|\,\xi - \eta'\,| - \Sigma\,|\,\xi - \eta\,|}{\Sigma\,|\,\eta - \eta'\,|} \qquad (3)$$

proposed by Gini (1914). Such a measure has advantages over Spearman's ρ, but not over Kendall's τ, in respect of the rapidity with which it approaches normality.

The sampling distribution of G (for $n = 2,\ 3,\ 4,\ 5,\ 6,\ 7$), from a population in which each possible ranking occurs equally frequently, has been already worked out for the untied case. In this paper we want to examine the problem, which frequently arise in practice, in which ties are present in the rankings (*).

2. — *Short method of determining G.*

We consider first the untied case. Suppose, for example, seven boys are ranked according to their ability in mathematics and in music

A Mathematics:	7	1	6	4	5	2	3
B Music:	3	2	1	7	6	5	4

Let us rearrange A in the natural order. For the second ranking we have the order

2 5 4 7 6 1 3

(*) For the determination of Gini's coefficient of rank correlation in the tied case, see: Salvemini (1939), Gini (1939), Grazia-Resi (1948).

For the untied case we may write

$$G = \frac{P - P'}{K}$$

where

$$P = \Sigma \, | \, \xi - \eta' \, | \, , \quad P' = \Sigma \, | \, \xi - \eta \, | \, ,$$

$$K = \frac{n^2}{2} \, , \, n \text{ even}$$

$$= \frac{n^2 - 1}{2} \, , \, n \text{ odd.}$$

We apply for the calculation of G a method as follows: We see that the first number, 2, has on its right 5 numbers which are greater. The second number, 5, has on its right 2 numbers which are greater and 1 number which is smaller on its left. The third number, 4, has on its right 2 numbers which are greater and 1 number on its left which is smaller, and so on. The contribution to P is therefore

$$P = 5 + 1 + 1 + 3 + 3 + 1 + 2 = 16 \, .$$

For the sum P' we can proceed inversely, as follows: The first number, 2, has on its right 1 number which is smaller. The second, 5, has on its right 3 numbers which are smaller. The third, 4, has on its right 2 numbers which are smaller and 1 number on its left which is greater, and so on. The contribution to P' is

$$P' = 1 + 3 + 1 + 3 + 1 + 5 + 4 = 18 \, .$$

The value of Gini's coefficient is given by

$$G = \frac{16 - 18}{\frac{49 - 1}{2}} = - \, 0{,}0833.$$

In general, if both rankings are in the order (perfect agreement):

$$1 \quad 2 \quad 3 \ldots . \, n \, ,$$

we get $P' = 0$,

$$P = (n - 1) + (n - 3) + \ldots + (n - 1) = K .$$

If the rankings are ranged in the inverse order (perfect disagreement), we have

$$P = 0 ,$$

$$P' = (n - 1) + (n - 3) + \ldots + (n - 1) = K .$$

3. — *Definition of G for rankings containing ties.*

We now examine the effect of ties on the calculation of Gini's coefficient. We shall adopt the well known "mid-rank method". For instance, if the observer ties the first and second member each is allotted the mean value 1,5 ; and if he ties the second, third and fourth each is allotted 3.

When ties are present, the above expression (3) requires some modification. For the tied case, Salvemini (1939) has given the formula

$$G_S = \frac{\Sigma \, | \, \xi - \eta' \, | - \Sigma \, | \, \xi - \eta \, |}{\Sigma \, | \, \xi^* - \eta'^* \, | - \Sigma \, | \, \xi^* - \eta^* \, |}$$

ξ^* and η^* being ξ and η disposed respectively in non-decreasing order ; moreover, the rankings η^* and η'^* are conjugate, so that $\eta^* + \eta'^* = n + 1$.

Gini (1939) has shown that, for the tied case, his rank correlation coefficient G may be conveniently calculated as follows

$$G_G = \frac{1}{2} (G_M + G_m) ,$$

where : the rank correlation coefficients G_M , G_m have been obtained by ranking the similar members in the sense of perfect agreement and disagreement respectively.

For the tied case, we consider the formula

$$C' = \frac{V (\xi + \eta) - V (\xi + \eta')}{\sqrt{V (2 \xi) \cdot V (2 \eta)}} . \qquad (4)$$

This general expression includes several coefficients of rank correlation as particular cases which arise when opportune measures of dispersion are adopted. For instance, if we assume, as a convenient measure of dispersion, the mean deviation about the mean, we find from (4)

$$G_A = \frac{\Sigma\,|\,\xi - \eta'\,| - \Sigma\,|\,\xi - \eta\,|}{\sqrt{\Sigma\,|\,\xi - \xi'\,| \cdot \Sigma\,|\,\eta - \eta'\,|}}\,.$$

It may be written more simply as

$$G_A = \frac{P - P'}{\sqrt{K \cdot K'}}\,.$$

For example, eight students have been arranged according to their merit in mathematics and in statistics :

Mathematics :	18	21	23	24	24	27	27	27
Statistics :	12	18	18	19	20	27	27	30

It follows the two rankings of 8 :

A :	1	2	3	4,5	4,5	7	7	7
B :	1	2,5	2,5	4	5	6,5	6,5	8

We may apply a short method of determining G_A as for the untied case. For the first number of B, the contribution to P is 7. The second and third number of B are tied ; the second and third number of A are untied, we shall allot $\frac{1}{2}$.

We see that the second number of B has on its right 5 numbers which are greater and 1 number on its left which is smaller, and so on. The sixth and seventh number of A are tied and so for B, we shall allot zero. Therefore, the sixth and seventh contribute to P for the same quantity 4,5. The sum P is then given by

$$P = 7 + 4{,}5 + 3{,}5 + 0{,}5 + 0{,}5 + 4{,}5 + 4{,}5 + 6 = 31\,.$$

The sum P' is

$$P' = 0 + 0{,}5 + 0{,}5 + 0{,}5 + 0{,}5 + 0{,}5 + 0{,}5 + 1 = 4.$$

For K and K' we can proceed as follows : the first number of A has, on its right, 7 numbers which are greater. The second has 6 numbers which are greater on its right and 1 number which is smaller on its left. The contribution to K is therefore 5. For the third number the contribution to K is 3 ; and so on. The sum K is

$$K = 7 + 5 + 3 + 0 + 0 + 5 + 5 + 5 = 30 .$$

The first number of B has, on its right, 7 numbers which are greater. The second has 5 numbers which are greater on its right and 1 which is smaller on its left. The contribution to K' is 4 ; and so on. The sum K' is

$$K' = 7 + 4 + 4 + 1 + 1 + 4 + 4 + 7 = 32.$$

The value of G_A, in this tied case, is given by

$$G_A = \frac{31 - 4}{\sqrt{30 \times 32}} = 0{,}872 .$$

4. — *Applications to particular cases.*

Consider the problem where one ranking is in the natural order and the other has the first $n - 1$ numbers tied :

$$\begin{array}{cccccc} 1 & 2 & 3 & \ldots . & n-1 & n \\ n/2 & n/2 & n/2 & \ldots . & n/2 & n . \end{array}$$

Omitting all algebraical reductions, we find

$$G_G = \frac{3n - 2}{n^2} \quad , \quad G_A = \frac{1}{\sqrt{n - 1}} , \; n \text{ even}$$

$$= \frac{3}{n + 1} \quad , \quad = \frac{n}{(n - 1)\sqrt{n + 1}} , \; n \text{ odd.}$$

For $n = 2$, G_G and G_A are equal to 1 ; in this case there is, in fact, complete agreement between the two rankings. For large n, G_G and G_A are nearly zero.

Suppose, for instance, an other particular problem where one ranking is in the natural order and the other has $n - 2$ numbers tied with ranks each $\frac{n+1}{2}$:

1	2	3	....	$n-1$	n
1	$\frac{n+1}{2}$	$\frac{n+1}{2}$	....	$\frac{n+1}{2}$	n.

We obtain

$$G_G = \frac{4(n-1)}{n^2}, \qquad G_A = \frac{2\sqrt{n-1}}{n}, \quad n \text{ even}$$

$$= \frac{4}{n+1}, \qquad = \frac{2}{\sqrt{n+1}}, \quad n \text{ odd.}$$

For $n = 2$ and $n = 3$ there is perfect agreement ; in fact, for these two special cases, $G_G = G_A = 1$. Moreover, G_G and G_A tend to zero as $n \to \infty$.

5. — *Methods of determining G_A when the variates are dichotomised.*

Suppose we have, as in psychology, the relationship between two rankings, one of which consists of a dichotomy with x and $y = n - x$ members in the two classes, as follows (see : Kendall, 1948)

Rank :	1	2	4	7	8	9	11	13	3	5	6	10	12	14	15
Sex :	b	b	b	b	b	b	b	b	g	g	g	g	g	g	g

In this example, there are 8 boys (b) and 7 girls (g) according to merit in an examination. We may write the rankings

A :	1	2	4	7	8	9	11	13	3	5	6	10	12	14	15
B :	4,5	4,5	4,5	4,5	4,5	4,5	4,5	4,5	12	12	12	12	12	12	12

For the calculation of G_A, we see that the first number of B has, on its right, 7 numbers which are greater ; and so for the first 8 numbers. The ninth number of B has 8 numbers which are smaller on its left ; and so for the last 7 numbers.

The contribution to K is therefore $7 \times 8 + 8 \times 7 = 112$.
In general, we have the formulae

$$G_A = \frac{D}{n\sqrt{x\,y}}, \quad n \quad even$$

$$= \frac{D}{\sqrt{x\,y\,(n^2 - 1)}}, \quad n \quad odd.$$

In our present instance

$$D = P - P' = 18, \quad x = 8, \quad y = 7, \quad n = 15;$$

and hence

$$G_G = 0{,}205; \quad G_S = 0{,}214; \quad G_A = 0{,}161.$$

In the next pages we shall discuss about the significance of observed rank correlations in the sense of the statistical theory of sampling.

Let us now consider the extreme case when both rankings consist of dichotomies : one into x and $y = n - x$ members and the other into p and $q = n - p$ members. In such a case, G_A is given by the simple form

$$G_A = \frac{D}{2\sqrt{x\,y\,p\,q}}.$$

When both variates are so tied, there is an analogous formula for the coefficient τ (see : Kendall, 1948) defined by

$$\tau = \frac{S}{\sqrt{x\,y\,p\,q}},$$

where $S = \overline{cov}\,(\xi, n)$.

It is well known that the Kendall's τ is an useful measure of association in a 2×2 contingency table. We shall see that G_A is also a convenient coefficient of the same kind.

A frequency distribution presented as a double dichotomy is shown, for example, in the following table

(η) \ (ξ)	Boys	Girls	Total
Using the theme of violence..	a	b	p
Not using the theme of violence	c	d	q
Total	x	y	n

We may express the value of D in terms of a, b, c, d. In this case, as may easily be demonstrated, we have

$$D = 2S + \frac{H}{2} \tag{5}$$

where

$$S = ad - bc,$$

$$H = |a - d|(b + c) - |b - c|(a + d);$$

and hence $D \gtrless 2S$, or $G_A \gtrless \tau$, if $H \gtrless 0$ respectively.

Let us now consider a practical problem. The results of an experiment on protection against a certain disease are illustrated in a 2×2 contingency table, as follows :

Group	No. of animals		Total
	Surviving (28)	Dead (78)	
Treated (30,5)	48	12	60
Untreated (80,5)	7	33	40
Total	55	45	100

From these results, we find

$$S = 48 \times 33 - 12 \times 7 = 1.500 ,$$

$$H = 15 \times 19 - 5 \times 81 = -120 ,$$

$$D = 3.000 - 60 = 2.940 .$$

It is useful to remark that the variables

$$28\left(=\frac{x+1}{2}\right),\ 78\left(=x+\frac{y+1}{2}\right),\ 30{,}5\left(=\frac{p+1}{2}\right),\ 80{,}5\left(=p+\frac{q+1}{2}\right),$$

do not occur in the calculation of D.

From this experiment, $H < 0$; so that $G_A < \tau$. We have, in fact, the values

$$G_A = 0{,}603 , \quad \tau = 0{,}615 .$$

6. — *The variance of D for the untied case.*

To test whether an observed value of D is "significant" or whether there is "conformity" in the Gini (1939) and Pompilj (1948) sense, it is necessary to work out the sampling distribution of D (or of G) in the null hypothesis where no parental correlation exists. Permuting the ranks in all possible ways, we get $n!$ values of D by supposing one ranking fixed.

The exact distribution of D has been obtained by Savorgnan (1915) and Salvemini (1951) for lower values of n ($n = 2, 3, \ldots, 7$). Salvemini (1951) has also furnished a table for the probability that D attains or exceeds a specified value, for $n = 4, 5, 6, 7$.

It is well known that, if the sample is sufficiently large, the sampling distribution of D, of all possible random samples of size n, follows the normal curve, defined by

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{X^2}{2\sigma^2}}$$

σ being the standard deviation.

If n is restricted to odd values, the varience of D is given by the formula (see : Amato, 1954)

$$var\ D = \frac{(n+1)(n^2+3)}{6} . \qquad (6)$$

For even values of n, the variance of D belongs to the interval

$$\{\varphi(n-1),\ \varphi(n+1)\}$$

where

$$\varphi(n) = \frac{(n+1)(n^2+3)}{6} .$$

7. — *Tendency of G to normality.*

Hotelling and Pabst (1936) have proved that the moments of ρ tend to those of the normal distribution as $n \to \infty$:

$$\beta_\nu(\rho) = \frac{\mu_{2\nu}(\rho)}{\mu_2^\nu(\rho)} \to \frac{(2\nu)!}{2^\nu \nu!} .$$

It follows, from the second limit theorem, that the distribution of ρ tends to normality.

Analogously, Kendall (1938) has demonstrated, as n tends to infinity, that

$$\beta_\nu(\tau) = \frac{\mu_{2\nu}(\tau)}{\mu_2^\nu(\tau)} \to \frac{(2\nu)!}{2^\nu \nu!} ;$$

and hence the sampling distribution of τ tends to normality as n increases. The tendency of τ to normality is more rapid than that of ρ.

For Gini's coefficient of rank correlation we see that the sampling distribution of G tends, as for ρ and τ, to normality. We have shown, for $n > 3$ (see : Amato, 1954), that

$$\beta_\nu(\rho) < \beta_\nu(G) < \beta_\nu(\tau)$$

where

$$\beta_\nu(G) = \frac{\mu_{2\nu}(G)}{\mu_2^\nu(G)} .$$

Thus, when n tends to infinity, $\beta_\nu(\rho)$ and $\beta_\nu(\tau)$ tend to a common limit

$$\frac{(2\nu)!}{2^\nu \nu!};$$

it follows that $\beta_\nu(G)$ tends to the same limit as $n \to \infty$.

The sampling distribution of G tends to normality more slowly than that of τ and more rapidly than that of ρ. For $n > 10$ we may use the table of areas under the normal curve.

8. — *The variance of D for the tied case.*

As for the untied case, the sampling distribution of D tends to normality as the sample size increases, unless the extent of the ties is very large.

When the size of sample is small and ties are present in the rankings, the variance of D requires some modification. We consider the case where only one ranking contains ties.

For instance, if one ranking is untied and the other has $t = n - 2$ members tied with ranks each $\frac{n+1}{2}$ and two members ranked 1 and n, the sampling distribution of D is given by

Value of D	Frequency
$-2(n-1)$	1
$-2(n-2)$	2
.	.
.	.
.	.
.	.
-2	$n-1$
2	$n-1$
.	.
.	.
.	.
$2(n-2)$	2
$2(n-1)$	1
	$n(n-1)$

In this case there are

$$\frac{n\,!}{(n-2)\,!} = n\,(n-1)$$

possible arrangements. The variance of this distribution, as can easily be proved, is

$$var\ D = \frac{2\,n\,(n+1)}{3}\,. \qquad (7)$$

Similarly, we may obtain some analogous expressions for the variance of D in the cases where one ranking is always untied and the other has $t = n-4$, $n-6$, $n-8$, ... members tied with ranks each $\frac{n+1}{2}$. In such a case, the formula (6) for the variance of D, abtained by correlating one ranking with all $n\,!$ possible arrangements of the other, requires some modification. If the integers n and t are odd and in one ranking there is a tie of t numbers equal to $\frac{n+1}{2}$, the second moment od D is defined by

$$var\ D = \frac{(n+1)\,(n^2+3) - (t+3)\,(t^2-1)}{6} \qquad (8)$$

$$t \leq n-2\,.$$

For the prove of this result, we consider some particular cases in which there are 1, 3, 5, ... numbers equal to the mean value of the first n integers. For n restricted to odd values, we have the formulae

$$t = 1\,, \quad var\ D = \frac{1}{6}\,[(n+1)\,(n^2+3) - (0 \times 2 \times 4)]$$

$$t = 3\,, \quad var\ D = \frac{1}{6}\,[(n+1)\,(n^2+3) - (2 \times 4 \times 6)]$$

$$t = 5\,, \quad var\ D = \frac{1}{6}\,[(n+1)\,(n^2+3) - (4 \times 6 \times 8)]$$

....

....

It follows the general expression (*)

$$var\ D = \frac{1}{6}\left[(n+1)(n^2+3) - (t-1)(t+1)(t+3)\right].$$

This is the result given by (8).

In particular, for $t = n - 2$ we find from (8) the second moment of D given by (7).

For instance, the sampling distribution of D, for $n = 5$ and $t = 3$, is

Value of D	Frequency
−8	1
−6	2
−4	3
−2	4
2	4
4	3
6	2
8	1
	20

(*) In this case we find

$$\Sigma\,|\,\eta - \eta'\,|\begin{cases} = \dfrac{n^2}{2} - \dfrac{t^2}{2}; & n, t \quad even \\ = \dfrac{n^2-1}{2} - \dfrac{t^2-1}{2}; & n, t \quad odd. \end{cases}$$

We may obtain these formulae from

$$\frac{n^2}{2} - 2a^2, \quad n \quad even$$

$$\frac{n^2-1}{2} - 2a(a+1), \quad n \quad odd$$

(see : SALVEMINI, 1939) putting $a = \dfrac{t}{2}$ and $a = \dfrac{t-1}{2}$ respectively.

For $n = 7$ and $t = 3$, we have obtained the distribution

Value of D (*)	Frequency
0	90
2	72
4	76
6	60
8	47
10	40
12	39
14	24
16	12
18	4
20	1
Total (of whole distribution)	840

The actual distributions form the basis of table 1. Such a table shows the probability that D attains or exceeds a specified value.

TABLE 1.

D	$n = 5$ $t = 3$	$n = 7$ $t = 3$	$n = 7$ $t = 5$	$n = 9$ $t = 7$	$n = 11$ $t = 9$
2	0,500	0,446	0,500	0,500	0,500
4	0,300	0,361	0,357	0,389	0,409
6	0,150	0,270	0,238	0,292	0,327
8	0,050	0,199	0,143	0,208	0,255
10		0,143	0,071	0,139	0,191
12		0,095	0,024	0,083	0,136
14		0,049		0,042	0,091
16		0,020		0,014	0,055
18		0,006			0,027
20		0,001			0,009

For a sample of size 7 and $t = 3$, we find, for example, $D = 14$. From (8) we obtain

$$\sqrt{var\, D} = \sqrt{\frac{8 \times 52}{6} - \frac{6 \times 8}{6}} = 7{,}8315\,.$$

(*) We show only positive values of D, the negative values being given by symmetry.

With a "correction for continuity" (see : Kendall, 1948)

$$\frac{D - 1}{\sqrt{var\ D}} = \frac{13}{7{,}8315} = 1{,}66 .$$

Now, the probability of a deviation of 1,66 times the standard error (*) or greater in absolute value, is $2\,(1 - 0{,}9515) = 0{,}097$. The exact value from table 1 is $2 \times 0{,}049 = 0{,}098$. For D no corrected for continuity we should have found 0,073.

For moderate t and large n, we may use the asymptotic formula

$$var\ D \approx \frac{3}{2}\ var\ S \qquad (9)$$

where

$$var\ S = \frac{n(n-1)\,(2n+5) - \sum_t t\,(t-1)\,(2t+5)}{18} \qquad (10)$$

(see : Sillitto, 1947).

Supposing t fixed, we see from table 2 that the variance of D tends to $\frac{3}{2}\ var\ S$ as the sample size increases. For large odd n it rapidly approaches $\frac{3}{2}\ var\ S$ for $t = 3$. In this case, the effect of the tie on the calculation of the variance is evidently small.

For the prove of (9) we have from (8) and (10) for $t = 3$ and *odd* n

$$\frac{var\ D}{var\ S} = \frac{3\,(n+1)\,(n^2+3) - 48}{n\,(n-1)\,(2n+5) - 66};$$

and hence for large n

$$\frac{var\ D}{var\ S} \approx \frac{3}{2} .$$

(*) The standard deviation of the sampling distribution is called the standard error.

TABLE 2.

n	$t = 3$		$t = 5$		$t = 7$	
	var D	$\frac{3}{2}$ *var S*	*var D*	$\frac{3}{2}$ *var S*	*var D*	$\frac{3}{2}$ *var S*
5	20,00	19,50	—	—	—	—
7	61,33	61,01	37,33	41,51	—	—
9	132,00	132,50	108,00	113,00	60,00	71,50
11	240,00	242,00	216,00	222,50	168,00	181,00
13	393,33	397,50	369,33	378,00	321,33	336,50
15	600,00	607,00	576,00	587,50	528,00	546,00
17	868,00	878,50	844,00	859,00	796,00	817,50
..	....	....	....	....	....	....
31	5133,33	5187,00	5109,33	5167,50	5061,33	5126,00
..	....	....	....	....	....	....
91	127013,33	127622,00	126989,33	127602,50	126941,33	127561,00

For the intermediate problems, no simple method is available. But we can always apply the inequalities

$$var\ S < var\ D < \frac{(n+1)(n^2+3)}{6} \qquad (11)$$

where *var S* and *var D* must be regarded as variances for the tied case. The argument will be clear from the applications illustrated at the end of this paper.

9. — *The variance of D when the rankings are dichotomised.*

If one ranking is untied and the other is a dichotomy into $x = n - 1$ and $y = 1$, we find, without detailed proofs, that

$$var\ D = \frac{(n+1)(n+2)}{3}, \quad n\ even$$

$$= \frac{(n-1)(n^2+4n+6)}{3n}, \quad n\ odd.$$

When both rankings become dichotomies, it is well known the formula for the variance of S (see : Kendall, 1948) :

$$var\ S = \frac{x\,y\,p\,q}{n-1}. \qquad (12)$$

For the variance of D we have from (5)

$$var\ D = 4\ var\ S + \frac{1}{4}\ var\ H + 2\ cov\ (S\ ,\ H)\ . \qquad (13)$$

If S and H are independent, we find from (13) and (12)

$$var\ D = \frac{4\ x\ y\ p\ q}{n - 1} + \frac{1}{4}\ var\ H\ ;$$

and hence

$$var\ D \geq \frac{4\ x\ y\ p\ q}{n - 1}\ .$$

We have the equality

$$var\ D = \frac{4\ x\ y\ p\ q}{n - 1} \qquad (14)$$

only if H is a constant or equal to zero.

For a four-fold table as the following

(η) \ (ξ)			
	64	32	96
	28	56	84
	92	88	180

we find $H = 0$ and therefore

$$D = 2\ (64 \times 56 - 32 \times 28) = 5.376.$$

For the variance of D we have from (14)

$$var\ D = \frac{4 \times 92 \times 88 \times 96 \times 84}{179} = 1.458.908{,}2\ .$$

Thus, with a correction for continuity (see : Kendall, 1948)

$$\frac{D - n}{\sqrt{var\ D}} = \frac{5.376 - 180}{1.207{,}9} = 4{,}3 .$$

The probability that this is attained or exceeded in absolute value is 0,00002.

The value of our sample is 4,3 standard errors ; it could happen by chance in about 2 samples out of 100.000. From these results it fallows that for our sample there is no conformity in the Gini sense.

We shall set up, for the rank correlation problems, a 5 per cent conformity level corresponding to a deviation of 1,96 standard errors.

10. — *Two applications to practical cases.*

I - Hours and gross earnings of nonsupervisory employees in retail trade, 1951 (*) :

Type of store	Average weekly earning	Average weekly hour	Rank according to earning	Rank according to hour
General merchandise.....	37	36	1	1,5
Department and mail order..	44	38	3	3
Food and liquor	54	40	4	4
Auto and accessory	67	45	7	7
Apparel and accessory	42	36	2	1,5
Furniture and appliance	60	43	6	5
Lumber and hardware	59	44	5	6

The observed D is 21. From (11) we may write

$$\frac{D}{\sqrt{var\ D}} > \frac{D}{\sqrt{\frac{(n+1)(n^2+3)}{6}}} = \frac{21}{\sqrt{\frac{8 \times 52}{6}}} = 2{,}52 .$$

*) «Monthly Labor Review»,.....

The probability that 2,52 times the standard error is attained or exceeded in absolute value is about 0,01174. Our probability is lesser than this and therefore we incline to attribute no conformity. In other words, there is no conformity among the sample D and the hypothesis of independence.

II - Birth rates for Italian Regions, 1930 (see : *Compendio Statistico Italiano*, 1931 - Roma) and average number of persons per natural family, 1931 (see : L. Livi, 1940) :

Rank according to birth rate	Rank according to family size
1	2
2	1
3	14
4	7,5
5	10
6	15
7	5
8	17,5
9	17,5
10	16
11	3
12	10
13	10
14	13
15	12
16	7,5
17	5
18	5

The observed D is 8. From (10) we find

$$\sqrt{var\, S} = \sqrt{\frac{1}{18}(18\times 17\times 41 - 2\times 2\times 1\times 9 - 2\times 3\times 2\times 11)} = 26{,}22\,;$$

and from (11)

$$\frac{D}{\sqrt{var\, D}} < \frac{8}{26{,}22} = 0{,}31\,.$$

The probability is greater than 0,7566. This is very large, for that there is conformity among the value of D and the hy-

pothesis that the two variables, birth rate and family size, are independent. In this case, the observed deviation between the sample D and zero is considered due to chance.

REFERENCES

1. — Amato V. (1952-'53), *Sulla determinazione degli indici di cograduazione mediante gli indici di variabilità*, Monografia n. 6 della Soc. Italiana di Econ. Demografia e Stat., Roma.

2. — Amato V. (1954), *Un criterio per la determinazione di indici di cograduazione tra serie statistiche con ripetizioni*, « Statistica », n. 1, gennaio - marzo.

3. — Amato V. (1954), *Sulla distribuzione dell'indice di cograduazione del Gini*, in : Studi in onore di G. Pietra, « Statistica », n. 3, luglio. settembre.

4. — Gini C. (1914), *Di una misura delle relazioni tra le graduatorie di due caratteri*, in « Saggi monografici », Roma.

5. — Gini C. (1915-'16), *Indici di concordanza*, « Atti del R. Istituto Veneto di Scienze, Lettere ed Arti », Tomo LXXV.

6. — Gini C. (1939) : *Sulla determinazione dell'indice di cograduazione*, « Metron », Vol. XIII, N. 4.

7. — Gini C. (1939), *I pericoli della statistica*, « Atti della 1ª riunione scientifica della Soc. Italiana di Statistica », Pisa.

8. — Gini C. (1945-'46), *Intorno alle basi logiche e alla portata gnoseologica del metodo statistico*, « Statistica ».

9. — Grazia-Resi B. (1948), *Nuove ricerche sugli indici di cograduazione fra serie con termini uguali*, «Statistica », n. 4.

10. — Hotelling H. and Pabst M.R. (1936), *Rank correlation and tests of significance involving no assumption of normality*, « Annals of Mathematical Statistics».

11. — Kendall M.G. (1938), *A new measure of rank correlation*, « Biometrika ».

12. — Kendall M.G. (1948), *Rank correlation methods*, London, Griffin.

13. — Livi L. (1940), *Trattato di demografia*, Padova, Cedam.

14. — Pompilj G. (1948), *Teorie statistiche della significatività e conformità dei risultati sperimentali agli schemi teorici*, « Statistica ».

15. — Salvemini T. (1939), *L'indice di cograduazione del Gini nel caso di serie statistiche con ripetizioni*, « Metron », Vol. XIII, N. 4.

16. — Salvemini T. (1951), *Sui vari indici di cograduazione*, « Statistica ».

17. — Savorgnan F. (1915), *Sulla formazione dei valori dell'indice di cograduazione*, Cagliari.

18. — Sillitto G.P. (1947), *The distribution of Kendall's τ coefficient of rank correlation in rankings containing ties*, « Biometrika », Vol. XXXIV.

19. — Spearman C. (1904), *The proof and measurement of association between two things*, « Am. Journ. Psychol ».

STEFANIA GATTI

Su un limite a cui tendono alcune medie

È noto che, se la successione di positivi:

$$x_1, x_2, \ldots x_n, \ldots \tag{1}$$

ammette un limite finito e determinato λ, anche la media aritmetica e la geometrica dei primi n termini della (1) tendono, per n infinito, allo stesso limite λ [1]. L. Galvani [2] ha dimostrato che anche la media di potenze dei primi n termini della (1) data dall'espressione:

$$M_r = \sqrt[r]{\frac{\sum_{i=1}^{n} x_i^r}{n}} \tag{2}$$

se si intende di assumere per questa espressione la sola determinazione reale positiva, tende allo stesso limite λ, per qualunque valore di r che non faccia perdere significato alla (2), ossia per $r \neq 0$.

In questo lavoro dimostreremo che: *per n tendente all'infinito anche la media esponenziale E dei primi n termini della* (1) *data dall'espressione*:

$$c^E = \frac{c^{x_1} + c^{x_2} + \ldots c^{x_n}}{n} \tag{3}$$

---

[1] E. Cesaro, *Corso di analisi algebrica*, Bocca, Torino, 1894.

[2] L. Galvani, *Dei limiti a cui tendono alcune medie*, «Boll. Un. Mat. Ital.», 1927.

(dove c è un reale positivo diverso da zero e da 1) *tende a* λ *e, se* $\lambda \neq 0$, *anche la media biplana combinatoria potenziata del Gini* [1] *dei primi n termini della* (1) *data dall'espressione* :

$$B_{dq}^{cp} = \sqrt[cp-dq]{\frac{\binom{n}{c} \sum_{l=1} P_l^c (x_i^p)}{\binom{n}{d} \sum_{l=1} P_l^d (x_i^q)}} \tag{4}$$

(dove con $P_l^c (x_i^p)$ si indica il prodotto generico delle potenze p^{me} di c termini della (1) e la somma è estesa a tutte le $\binom{n}{c}$ combinazioni possibili c a c di detti termini, e simile significato ha $P_l^d (x_i^q)$ e dove inoltre p e q sono reali) *tende, per n all'infinito, allo stesso limite* λ *purchè sia* $c\,p - d\,q \neq 0$ *e si convenga di assumere il solo valore reale positivo dell'espressione* (4).

* * *

Per la dimostrazione di quanto ora asserito ci serviremo del seguente noto teorema [2] :

se, per n crescente all'infinito, la variabile b_n, *crescendo sempre, oltrepassa ogni limite, si ha* :

$$\lim_{n \to \infty} \frac{a_n}{b_n} = \lim_{n \to \infty} \frac{a_n - a_{n-1}}{b_n - b_{n-1}} \tag{5}$$

purchè esista il secondo membro.

Ciò posto dimostreremo in primo luogo che è :

$$\lim_{n \to \infty} \frac{\sum_{l=1}^{\binom{n}{c}} P_l^c (x_i^p)}{\binom{n}{c}} = \lambda^{cp} \tag{6}$$

---

[1] C. Gini, *Di una formula comprensiva delle medie*, «Metron» Vol. XIII, 1938.

[2] E. Cesaro, Op. cit.

Sia in un primo tempo $c = 2$ e p un reale qualsiasi. Il primo membro della (6) si potrà scrivere:

$$\lim_{n\to\infty} \frac{\Sigma\, x_i^p\, x_j^p}{\binom{n}{2}} \qquad (i<j)$$

ossia, in base alla (5):

$$\lim_{n\to\infty} \frac{x_1^p x_2^p + x_1^p x_3^p + \ldots x_1^p x_n^p + x_2^p x_3^p + \ldots + x_{n-1}^p x_n^p}{\binom{n}{2}} =$$

$$= \lim_{n\to\infty} \left\{ \frac{x_1^p x_2^p + x_1^p x_3^p + \ldots + x_1^p x_n^p + x_2^p x_3^p + \ldots x_{n-1}^p x_n^p}{\binom{n}{2} - \binom{n-1}{2}} - \right.$$

$$\left. - \frac{x_1^p x_2^p + x_1^p x_3^p \ldots x_1^p x_{n-1}^p + x_2^p x_3^p + \ldots + x_{n-2}^p x_{n-1}^p}{\binom{n}{2} - \binom{n-1}{2}} \right\} =$$

$$= \lim_{n\to\infty} \frac{x_1^p x_n^p + x_2^p x_n^p + \ldots x_{n-1}^p x_n^p}{n-1} = \lim_{n\to\infty} x_n^p \frac{x_1^p + x_2^p + \ldots + x_{n-1}^p}{n-1}$$

Osserviamo ora che se $\lim_{n\to\infty} x_n = \lambda$ è $\lim_{n\to\infty} x_n^p = \lambda^p$ e quindi, per quanto detto all'inizio di questo lavoro, la media aritmetica delle x_i^p avrà come limite:

$$\lim_{n\to\infty} \frac{x_1^p + x_2^p + \ldots + x_{n-1}^p}{n-1} = \lambda^p$$

Per cui si avrà in definitiva:

$$\lim_{n\to\infty} \frac{\Sigma\, x_i^p\, x_j^p}{\binom{n}{2}} = \lambda^{2p}$$

La (6) è quindi dimostrata per $c=2$. Supponiamo ora che essa sia vera per $c=k$; dimostreremo che, in tal caso, essa è vera anche per $c=k+1$. Infatti, per la (5), è:

$$\lim_{n\to\infty} \frac{\Sigma\, x_{a_1}^p x_{a_2}^p \dots x_{a_{k+1}}^p}{\binom{n}{k+1}} = \lim_{n\to\infty} \frac{x_n^p \Sigma\, x_{a_1}^p x_{a_2}^p \dots x_{a_k}^p}{\binom{n}{k+1} - \binom{n-1}{k+1}}$$

dove la somma a primo membro è estesa a tutte le possibili combinazioni di $k+1$ termini dei primi n termini della (1) e la somma a secondo membro è estesa a tutte le possibili combinazioni di k termini dei primi $n-1$ termini della (1).

La precedente è quindi uguale a:

$$\lim_{n\to\infty} x_n^p \cdot \lim_{n\to\infty} \frac{\Sigma\, x_{a_1}^p x_{a_2}^p \dots x_{a_k}^p}{\binom{n-1}{k}} = \lambda^p \cdot \lambda^{kp} = \lambda^{(k+1)p}$$

La (6) è così completamente dimostrata.

Ritorniamo ora alla formula (4) del Gini. Si ha:

$$\lim_{n\to\infty} \frac{\dfrac{\sum_{l=1}^{\binom{n}{c}} P_l^c(x_i^p)}{\binom{n}{c}}}{\dfrac{\sum_{l=1}^{\binom{n}{d}} P_l^d(x_i^q)}{\binom{n}{d}}} = \frac{\lim_{n\to\infty} \dfrac{\sum_{l=1}^{\binom{n}{c}} P_l^c(x_i^p)}{\binom{n}{c}}}{\lim_{n\to\infty} \dfrac{\sum_{l=1}^{\binom{n}{c}} P_l^c(x_i^q)}{\binom{n}{d}}} = \frac{\lambda^{cp}}{\lambda^{dq}} = \lambda^{cp-dq}$$

e quindi si ha subito:

$$\lim_{n\to\infty} B_{dq}^{cp} = \lambda$$

Calcoliamo ora il limite della media esponenziale.

Applicando la (5) si ha:

$$\lim_{n\to\infty} c^E = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} c^{x_i}}{n} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} c^{x_i} - \sum_{i=1}^{n-1} c^{x_i}}{n-(n-1)} = \lim_{n\to\infty} c^{x_n} = c^{\lambda}$$

da cui si ha:

$$\lim_{n\to\infty} E = \lambda$$

La proprietà quindi trovata per la media aritmetica, la geometrica e la media di potenze è valida anche per la media esponenziale e per ogni media che scaturisce dalla formula (4) del Gini per particolari valori di c, d, p, e q; in particolare avremo che allo stesso limite λ tendono, per n infinito, la media di somme di potenze e ogni media combinatoria.

SUMMARY

The authoress demonstrates that:
Let

$$x_1, x_2, \ldots x_n, \ldots \qquad (1)$$

be a sequence of positive numbers with $\lim_{n\to\infty} x_n = \lambda$ we have:

1) $\lim_{n\to\infty} C^E = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} C^{x_i} = C^{\lambda}$; hence $\lim_{n\to\infty} E = \lambda$;

(C is a real number different from 0 and 1)

and, if $\lambda \neq 0$:

2) $$\lim_{n\to\infty} B^{\frac{cp}{dq}} \lim_{n\to\infty} \sqrt[cp-dq]{\frac{\binom{n}{d} \sum_{l=1}^{\binom{n}{c}} P_l^c (x_i^p)}{\binom{n}{c} \sum_{l=1}^{\binom{n}{d}} P_l^d (x_i^q)}} = \lambda \; ; \; cp - dq \neq 0$$

where $P_i^c(x_i^p)$ is the generical product of the powers of order p of c numbers of (1) and the sum is taken over all the $\binom{n}{c}$ combinations of terms (1); $P_i^d(x_i^q)$ has an analogous meaning, moreover only the real values of B^{cp}_{dq} must be considered.

(p and q are real numbers).

E. J. GUMBEL and P. G. CARLSON

On the Asymptotic Covariance of the Sample Mean and Standard Deviation

SUMMARY

It is well known that the covariance of the sample mean $\bar{x}$ and standard deviation s is zero for the normal distribution, but its value for other populations seems to be unknown. Springer [3] has given the exact joint distribution of $\bar{x}$ and s in the form of a multiple integral which does not lead to simple solutions. In the following a general expression for the asymptotic covariance of these statistics for populations with all moments finite is given. Applications are made to the normal, logistic, and exponential distributions and pairs of uncorrelated linear functions of $\bar{x}$ and s are given.

1. *The Covariance of $\bar{x}$ and s.*

Let $\bar{x}$ and $s = \sqrt{m_2}$ be the mean and standard deviation of a sample of size n from a population for which all moments exist Let μ_i be the ith population central moment, $i = 1, 2, \ldots$. It will be shown that *

$$\text{Covariance} \quad (\bar{x}, s) = \mu_3/(2 n \sqrt{\mu_2}) + O\left(n^{-\frac{3}{2}}\right) \qquad (1.1)$$

*This was conjectured by C. DERMAN. (Columbia University, N. Y.)

The proof will be based on the values of the expectations and variances of sample moments and on the Schwarz Inequality for the orders of magnitude. Following Cramér [1], equ. 27.7, consider the identity

$$\begin{aligned}(\bar{x}-E\bar{x})-(\sqrt{m_2}-\sqrt{\mu_2}) &= (\bar{x}-E\bar{x})-(\sqrt{m_2}-E\sqrt{m_2})+O(n^{-1}) \\ &= (\bar{x}-E\bar{x})-\frac{(m_2-\mu_2)}{2\sqrt{\mu_2}}+\frac{(m_2-\mu_2)^2}{2\sqrt{\mu_2}(\sqrt{m_2}+\sqrt{\mu_2})^2}\end{aligned} \tag{1.2}$$

where E denotes the expectation. Squaring,

$$\begin{aligned}&[(\bar{x}-E\bar{x})-(\sqrt{m_2}-E\sqrt{m_2})]^2+2[(\bar{x}-E\bar{x})-(\sqrt{m_2}-E\sqrt{m_2})]\cdot \\ &\cdot O(n^{-1})+O(n^{-2})=(\bar{x}-E\bar{x})^2+\frac{(m_2-\mu_2)^2}{4\mu_2}+\frac{(m_2-\mu_2)^4}{4\mu_2(\sqrt{m_2}+\sqrt{\mu_2})^4}- \\ &-\frac{(\bar{x}-E\bar{x})(m_2-\mu_2)}{\sqrt{\mu_2}}+\frac{(\bar{x}-E\bar{x})(m_2-\mu_2)^2}{\sqrt{\mu_2}(\sqrt{m_2}+\sqrt{\mu_2})^2}-\frac{(m_2-\mu_2)^3}{2\mu_2(\sqrt{m_2}+\sqrt{\mu_2})^2}\end{aligned}$$

Taking expected values,

$$\begin{aligned}&Var(\bar{x}-\sqrt{m_2})+O(n^{-2})=E(\bar{x}-E\bar{x})^2+\frac{1}{4\mu_2}E(m_2-\mu_2)^2+ \\ &+\frac{1}{4\mu_2}E\left[\frac{(m_2-\mu_2)^4}{(\sqrt{m_2}+\sqrt{\mu_2})^4}\right]-\frac{1}{\sqrt{\mu_2}}[E(\bar{x}-E\bar{x})(m_2-\mu_2)]+ \\ &+\frac{1}{\sqrt{\mu_2}}E\left[\frac{(\bar{x}-E\bar{x})(m_2-\mu_2)^2}{(\sqrt{m_2}+\sqrt{\mu_2})^2}\right]-\frac{1}{2\mu_2}E\left[\frac{(m_2-\mu_2)^3}{(\sqrt{m_2}+\sqrt{\mu_2})^2}\right]\end{aligned} \tag{1.3}$$

The six factors on the right side of (1.3) are known and are given in textbooks, e.g. Cramér [1], equations 27.4 and 27.6.

$$E(\bar{x}-E\bar{x})^2=Var\,\bar{x}=\frac{\mu_2}{n} \tag{1.4}$$

$$\begin{aligned}&\frac{1}{4\mu_2}E(m_2-\mu_2)^2=\frac{1}{4\mu_2}E[(m_2-Em_2)+O(n^{-1})]^2= \\ &=\frac{1}{4\mu_2}E(m_2-Em_2)^2+O(n^{-2})=\frac{1}{4\mu_2}\left[\frac{\mu_4-\mu_2^2}{n}\right]+O(n^{-2})\end{aligned} \tag{1.5}$$

Now

$$\frac{(m_2-\mu_2)^4}{(\sqrt{m_2}+\sqrt{\mu_2})^4} < \frac{1}{\mu_2^2}(m_2-\mu_2)^4,$$

Hence,

$$\frac{1}{4\mu_2} E\left[\frac{(m_2-\mu_2)^4}{(\sqrt{m_2}+\sqrt{\mu_2})^4} < \frac{1}{4\mu_2^3} E(m_2-\mu_2)^4 = O(n^{-2}) \right. \tag{1.6}$$

$$\frac{1}{\sqrt{\mu_2}} E(\bar{x}-E\bar{x})(m_2-\mu_2) = \frac{1}{\sqrt{\mu_2}} E(\bar{x}-E\bar{x})(m_2-Em_2+O(n^{-1})) = \tag{1.7}$$

$$= \frac{1}{\sqrt{\mu_2}} E(\bar{x}-E\bar{x})(m_2-Em_2) = \frac{1}{\sqrt{\mu_2}} \cdot \frac{\mu_3}{n} + O(n^{-2})$$

It follows from (1.4), (1.6) and the Schwarz Inequality that the two remaining factors in (1.3) are of a lower order of magnitude. Indeed,

$$\left| E\left[(\bar{x}-E\bar{x})\left(\frac{m_2-\mu_2}{\sqrt{m_2}+\sqrt{\mu_2}}\right)^2\right]\right| \leq \tag{1.8}$$

$$\leq \left[E(\bar{x}-E\bar{x})^2 E\left(\frac{m_2-\mu_2}{\sqrt{m_2}+\sqrt{\mu_2}}\right)^4\right]^{\frac{1}{2}} = O\left(n^{-\frac{3}{2}}\right),$$

and from (1.5) and (1.6),

$$\left| E\left[\frac{(m_2-\mu_2)^3}{(\sqrt{m_2}+\sqrt{\mu_2})^2}\right]\right| \leq \left[E(m_2-\mu_2)^2 E\left(\frac{(m_2-\mu_2)^4}{(\sqrt{m_2}+\sqrt{\mu_2})^4}\right)\right]^{\frac{1}{2}} = O\left(n^{-\frac{3}{2}}\right) \tag{1.9}$$

Combining (1.4) thru (1.9) in (1.3)

$$Var(\bar{x}-\sqrt{m_2}) = \frac{\mu_2}{n} + \frac{1}{4\mu_2}\left[\frac{\mu_4-\mu_2^2}{n}\right] - \frac{1}{n}\frac{\mu_3}{\sqrt{\mu_2}} + O\left(n^{-\frac{3}{2}}\right) \tag{1.10}$$

Expanding the left side of (1.10) and making use of Cramér [1], 27.7,

$$\frac{\mu_2}{n} - 2\,Cov(\bar{x}, \sqrt{m_2}) + \frac{\mu_4-\mu_2^2}{4n\mu_2} = \frac{\mu_2}{n} - \frac{\mu_3}{n\sqrt{\mu_2}} + \frac{\mu_4-\mu_2^2}{4n\mu_2} + O\left(n^{-\frac{3}{2}}\right)$$

from which (1.1) follows.

An easy generalization of this result is

$$Var\,(a\bar{x}+b\,s)=a^2\frac{\mu_2}{n}+\frac{a\,b}{n}\frac{\mu_3}{\sqrt{\mu_2}}+b^2\frac{\mu_4-\mu_2^2}{4\,n\,\mu_2}+O\left(n^{-\frac{3}{2}}\right) \quad (1.11)$$

and for symmetrical distributions

$$Var\,(a\bar{x}+b\,s)=a^2\frac{\mu_2}{n}+b^2\frac{\mu_4-\mu_2^2}{4\,n\,\mu_2}+O\left(n^{-\frac{3}{2}}\right) \quad (1.11')$$

In the following paragraphs terms of $O\left(n^{-\frac{3}{2}}\right)$ and lower will be omitted.

2. *Uncorrelated Pairs of Statistics.*

It is asked whether there exist pairs of linear combinations of $\bar{x}$ and s, say $a\bar{x}+b\,s$, $c\bar{x}+d\,s$ which, asymptotically, have 1) proportional variances and 2) covariance zero. For the first condition let

$$Var\,(c\bar{x}+d\,s)=H^2,\quad Var\,(a\bar{x}+b\,s)=k^2\,H^2 \quad (2.1)$$

From (1.1), (1.11), and (2.1) it follows that

$$(ab+c\,d\,k^2)\frac{\mu_3}{n\sqrt{\mu_2}}=(k^2\,c^2-a^2)\frac{\mu_2}{n}+(k^2\,d^2-b^2)\frac{\mu_4-\mu_2^2}{4\,n\,\mu_2} \quad (2.2)$$

From the second condition,

$$cov\,(a\bar{x}+b\,s,\,c\bar{x}+d\,s)=a\,c\,Var\,\bar{x}+(a\,d+b\,c)\,cov\,(\bar{x},\,s)+bd\,Var\,s=0$$

or

$$a\,c\frac{\mu_2}{n}+(a\,d+b\,c)\frac{\mu_3}{2\,n\sqrt{\mu_2}}+b\,d\frac{\mu_4-\mu_2^2}{4\,n\,\mu_2}=0 \quad (2.3)$$

Solving (2.2) and (2.3) simultaneously gives the required values for a and b.

$$a=\pm\frac{k\,[2\,c\,\beta_1+d\,(\beta_2+2)]}{2\sqrt{(\beta_2+2)-\beta_1^2}},\qquad b=\mp\frac{k\,[2\,c+d\,\beta_1]}{\sqrt{(\beta_2+2)-\beta_1^2}} \quad (2.4)$$

where β_1 and β_2 are defined by $\beta_1 = \frac{\mu_3}{\mu_2^{3/2}}$, $\beta_2 = \frac{\mu_4}{\mu_2^2} - 3$. It has been shown by J. E. Wilkins [4], that $(\beta_2 + 2) - \beta_1^2 \geq 0$. Hence, for given c and d, (2.4) gives a and b satisfying the desired conditions.

For symmetrical distributions $\beta_1 = 0$ and

$$a = \pm \frac{k\,d\sqrt{\beta_2 + 2}}{2}, \quad b = \mp \frac{2\,k\,c}{\sqrt{\beta_2 + 2}} \tag{2.5}$$

is the required pair of factors which fulfill the two conditions.

3. *Application to Specific Distributions.*

For the normal distribution, $\beta_1 = \beta_2 = 0$. From (2.5), $c\bar{x} + d\,s$ and $\pm (d/\sqrt{2})\,\bar{x} \mp \sqrt{2}\,c\,s$ are a pair of uncorrelated statistics having equal variances. If, for example, $c = 1/\sqrt{2}$ and $d = 1$, then $(1/\sqrt{2})\,\bar{x} + s$ and $(1/\sqrt{2})\,\bar{x} - s$ are a specific pair of uncorrelated statistics having equal variances. Also, if $c = 0$, $d = 1$, and $k = \sqrt{2}$ in (2.5), $\bar{x}$ and s are uncorrelated and $Var\ \bar{x} = 2\ Var\ s$ as is well known.

For the logistic distribution

$$f(x) = \frac{\alpha\, e^{-\alpha(x-u)}}{(1 + e^{-\alpha(x-u)})^2}, \qquad -\infty < x < \infty \tag{3.1}$$

$\mu_2 = \pi^2/3\,\alpha^2$, $\beta_1 = 0$, and $\beta_2 = 1.2$ as shown in [2]. Hence from (2.5) $c\bar{x} + d\,s$ and $\pm (\sqrt{0.8})\,d\,\bar{x} \mp (1/\sqrt{0.8})\,c\,s$ are a pair of uncorrelated statistics having equal variances. If, for example, $c = \sqrt{0.8}$ and $d = 1$, then $\sqrt{0.8}\,\bar{x} + s$ and $\sqrt{0.8}\,\bar{x} - s$ are uncorrelated statistics having variances both equal to $1.6\,\pi^2/3\,\alpha^2 n$.

Consider a limited exponential distribution with parameters α and ϵ,

$$\begin{aligned} f(x, \alpha, \epsilon) &= \alpha\, e^{-\alpha(x-\epsilon)}, \quad x \geq \epsilon \\ &= 0 \qquad\qquad , \quad x < \epsilon \end{aligned} \tag{3.2}$$

Simple calculations show that

$$E\,x = \epsilon + \frac{1}{\alpha}\,,\quad E\,x^2 = \epsilon^2 + 2\,\frac{\epsilon}{\alpha} + 2\,\frac{1}{\alpha^2}\,,\quad Var\,x = \frac{1}{\alpha^2} = \mu_2 \tag{3.3}$$

$$\mu_4 = 9\,\mu_2^2\,,\quad \mu_3 = 2\,\mu_2\sqrt{\mu_2}\,,\quad \beta_1 = 2\,,\quad \beta_2 = 6 \tag{3.4}$$

Hence, for exponential distributions (2.4) becomes

$$a = \pm\, k\,(c + 2\,d)\,,\, b = \mp\, k\,(c + d) \tag{3.5}$$

and $c\,\bar{x} + d\,s$ and $\pm\,(c + 2\,d)\,\bar{x} \mp (c + d)\,s$ are uncorrelated statistics having equal variances. From (3.3) $\hat{\epsilon} = \bar{x} - s$ ($k = 1$, $c = 1$, $d = -1$) is an estimator of the lower limit ϵ. This statistic can, of course, be used only when $\hat{\epsilon}$ is smaller than the smallest observation. It follows from (1.11) that

$$Var\,\hat{\epsilon} = \frac{\mu_2}{n} + O\,(n^{-3/2}) \tag{3.6}$$

equals asymptotically the variance of the sample mean $\bar{x}$.

By (3.5) $\hat{\epsilon}$ and $\bar{x}$ are uncorrelated. Thus, at least from the viewpoint of asymptotic sampling variance, $\hat{\epsilon}$ is as reliable an estimator of the lower limit ϵ as $\bar{x}$ is of the population mean. It follows also that

$$\hat{\epsilon}^* = \bar{x} - \sqrt{\frac{n}{n-1}}\,s \tag{3.7}$$

which is an unbiased estimator of ϵ has the same asymptotic properties as $\hat{\epsilon}$. Further, $\bar{x}$ and $\hat{\epsilon}$ are asymptotically normally and independently distributed with equal variances as follows from Cramér [1], 28.4. This property may prove helpful in the estimation of the lower limit of an exponential distribution.

For example, consider eight observations reported by Sukhatme in Table 1 of [4] on the lengths of intervals between successive telephone calls. The ordered observations are 1, 3, 3, 15, 25, 33, 39, 70 in half-minute units. The mean is $\bar{x} = 23.6$ units,

and the standard deviations $S = 23.6$ units. Hence, the estimate of the lower limit using [3.7] is zero, which is less than the smallest sample value.

Conclusion :

There is derived the asymptotic covariance of the sample mean and standard deviation. This leads to the errors of estimate for two-parameter distributions, where the parameters are estmated by linear functions of the mean and standard deviations.

REFERENCES

[1] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[2] E. J. Gumbel, *Ranges and Midranges*, « Annals of Mathematical Statistics », Vol. 15 (1944), p. 414.

[3] M. D. Springer, *Joint Sampling Distribution of the Mean and Standard Deviation for Probability Density Functions of Doubly Infinite Range*, « Annals of Mathematical Statistics », Vol. 24 (1953), p. 118.

[4] P. V. Suktatme, *On the analysis of k samples from exponential populations with especial reference to the problem of random intervals*, « Statistical Research Memoirs », edited by Neyman and Pearson, p. 95 1936.

[5] J. E. Wilkins, *A Note on Skewness and Kurtosis*, « Annals of Mathematical Statistics », Vol. 15 (1944), p. 333.

CARLO BENEDETTI

Sulla rappresentabilità di una distribuzione binomiale mediante una distribuzione *B* e viceversa [1]

In questo lavoro, data una distribuzione binomiale:

$$f_b(x) = \binom{n}{n\,x} p^{nx}\, q^{n(1-x)}$$

con x variabile tra 0 e 1, estremi inclusi e percorrente i valori $x = 0, 1/n, 2/n, \ldots \frac{n-1}{n}, 1$; e tenendo n finito, viene considerato un semplice criterio per rappresentare soddisfacientemente tale distribuzione con una legge del tipo:

$$f_B(x) = \frac{1}{B(l, m)}\, x^{l-1}\,(1-x)^{m-1} \quad (0 \leq x \leq 1;\, l > 1,\, m > 1)$$

dove:

$$B(l, m) = \int_0^1 x^{l-1}\,(1-x)^{m-1}\, dx = \frac{\Gamma(l)\,\Gamma(m)}{\Gamma(l+m)}$$

cioè con la nota distribuzione B (l ed m reali). Inoltre in relazione a queste due distribuzioni vengono esposte alcune considerazioni sulla distribuzione delle probabilità *a priori* nella formula di Bayes.

* * *

Anzitutto dobbiamo notare che mentre la $f_B(x)$ rappresenta una densità, cioè la lunghezza dell'ordinata nel punto x, la $f_b(x)$ per un certo $x = \frac{i}{n}$ rappresenta l'area del rettangolino di base

[1] Questo lavoro è stato oggetto di una comunicazione al Seminario di Statistica della Facoltà di Scienze Statistiche, Demografiche ed Attuariali della Università di Roma il 24 Giugno 1954.

$1/n$ e altezza $f_b(x)\,n$, quindi per comparare le due curve possiamo comparare con $f_b(x)$ una delle tre espressioni:

$$\beta_1 = \int_{x-\frac{1}{n}}^{x} f_B(t)\,dt\left(\frac{1}{n} \leq x \leq 1\right); \quad \beta_2 = \int_{x-\frac{1}{2n}}^{x+\frac{1}{2n}} f_B(t)\,dt\left(\frac{1}{2n} \leq x \leq 1 - \frac{1}{2n}\right);$$

$$\beta_3 = \int_{x}^{x+\frac{1}{n}} f_B(t)\,dt\left(0 \leq x \leq 1 - \frac{1}{n}\right);$$

agli effetti delle applicazioni numeriche allegate abbiamo scelto:

$$\beta_2 \cong \frac{1}{n} f_B(x) \rightarrow \quad f_b(x)$$

Appurato questo, la soluzione assai semplice e spontanea che si offre per rappresentare con una certa approssimazione la $f_b(x)$ con la $f_B(x)$ e viceversa, è quella di determinare i parametri incogniti che sono sempre due: p, n oppure l, m mediante la condizione che le due curve abbiano la stessa media e la stessa varianza, e cioè siccome abbiamo:

$$Med\,[x, f_b(x)] = p = \sum_{x=0}^{1} x \binom{n}{n\,x} p^{nx} q^{n(1-x)}$$

$$Med\,[x, f_B(x)] = \frac{l}{l+m} = \frac{1}{B(l, m)} \int_0^1 x^l (1-x)^{m-1}\,dx$$

$$Var\,[x, f_b(x)] = \frac{p\,q}{n} = \sum_{x=0}^{1} (x-p)^2 \binom{n}{n\,x} p^{nx} q^{n(1-x)}$$

$$Var\,[x, f_B(x)] = \frac{l\,m}{(l+m)^2 (l+m+1)} =$$

$$= \frac{1}{B(l, m)} \int_0^1 \left(x - \frac{l}{l+m}\right)^2 x^{l-1} (1-x)^{m-1}\,dx$$

avremo il seguente sistema di due equazioni con due incognite:

$$\begin{cases} p = \dfrac{l}{l+m} \\ \dfrac{1}{n}\, p\,(1-p) = \dfrac{l\,m}{(l+m)^2\,(l+m+1)} \end{cases}$$

che risolto dà le seguenti relazioni tra l, m e p, n

$$l = (n-1)\,p \quad ; \quad p = \frac{l}{n-1} \quad ; \quad n = l+m+1\ (^1) \qquad (:)$$

$$m = (n-1)\,q \qquad q = \frac{m}{n-1} \qquad n > \frac{1+p}{p}\,,\ n > \frac{1+q}{q}$$

che consentono alle due curve $f_b(x)$ ed $f_B(x)$ di avere la stessa media e la stessa varianza.

È facile vedere che le (:) consentono anche la pratica coincidenza delle mode, e tale coincidenza diviene sempre più effettiva al crescere di n. Infatti la moda di $f_B(x)$ è (uguagliando a zero la derivata prima e risolvendo rispetto ad x):

$$x_{moda} = p + \frac{2\,p}{n-3} - \frac{1}{n-3} \simeq p \text{ per } n \text{ abbastanza}$$

grande e la moda di $f_b(x)$ è soggetta, come è noto, alla limitazione:

$$p - \frac{q}{n} \leq x_{moda} \leq p + \frac{p}{n}$$

che per n abbastanza grande significa che $x_{moda} \simeq p$.

Evidentemente altre soluzioni sono possibili adoperando anzichè la media e la varianza, altri momenti di grado più elevato.

---

(1) Se $l+m+1$ non è intero (quando l ed m sono dati) si prenderà l'intero più vicino.

Come è ovvio, nel caso di n finito, non ha senso parlare di perfetta identità tra le due distribuzioni in quanto la beta è continua per ogni $0 < x < 1$ mentre la binomiale ha solo valori per $n\,x$ intero.

Invece nel caso di n tendente all'infinito, caso però che in questo lavoro non interessa, e ponendo in accordo al nostro criterio

$$l = (n - 1)\,p\,;\quad m = (n - 1)\,q$$

dove p e q rimangono fissati, basta ricordare [1] che posto

$$\xi = \frac{x - p}{\sqrt{\frac{p\,q}{n}}}$$

abbiamo per la distribuzione-beta espressa mediante la variabile ξ

$$^{s}f_B(\xi) = \sqrt{\frac{p\,q}{n}}\, f_B\left(\xi\sqrt{\frac{p\,q}{n}} + p\right) \to \frac{1}{\sqrt{2\,\pi}}\, e^{-\frac{\xi^2}{2}}\,;$$

$$\int_{-\infty}^{t} {}^{s}f_B(\xi)\,d\,\xi \to \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{t} e^{-\frac{\xi^2}{2}}\,d\,\xi$$

in ciascuno dei suoi punti di continuità.

Così le distribuzioni binomiali e beta sotto le suddette condizioni trovano per $n \to \infty$ una comune rappresentabilità nella legge normale.

Mediante le relazioni trovate tra i parametri p, n ed l, m sono state fatte tre applicazioni i cui risultati, riportati nella tabella e nel grafico allegato, sono molto soddisfacienti.

---

[1] H. Cramèr, *Mathematical Methods of Statistics*, Princeton University Press. 1954, pag. 252 ; per rientrare nel caso trattato da questo Autore basta ovviamente porre $n - 1 = n'$.

I risultati soddisfacienti ottenuti autorizzano le seguenti considerazioni già avanzate dal Gini (1) relative all'ipotesi di Gini (2) sulla distribuzione delle probabilità a priori nella formula di Bayes.

Se in un campione di dimensione n tratto bernoullianamente da una certa massa di dimensione ν abbiamo riscontrato la frazione p di elementi difettosi e vogliamo risalire con la formula di Bayes alla probabilità di una certa frazione di elementi difettosi nella massa dobbiamo introdurre alcune ipotesi sulla distribuzione delle probabilità a priori della composizione della massa

Il Gini osserva che l'ipotesi che tutte le possibili composizioni della massa siano equiprobabili corrisponde all'ipotesi che tutte le *combinazioni* dei casi favorevoli e contrari ad ogni composizione della massa siano equiprobabili. Questa ipotesi, d'altra parte, non è giustificabile anche nell'assoluta ignoranza circa le probabilità a priori suddette.

Se, malgrado la completa assenza di conoscenze relative a tali probabilità, siamo costretti a prendere ugualmente una decisione la sola ipotesi attendibile sarebbe che fossero equiprobabili non tutte le *combinazioni* ma tutte le *disposizioni* dei casi favorevoli o contrari ad ogni composizione della massa poichè tutte queste disposizioni sono tutte ugualmente immaginabili ; ora ciò non porta ad una equidistribuzione delle combinazioni e quindi delle probabilità a priori, ma ad una equidistribuzione delle disposizioni e quindi ad una distribuzione binomiale.

Se non chè, dal punto di vista pratico, la distribuzione binomiale, quale distribuzione delle probabilità a priori, appesantisce enormemente la formula di Bayes, rendendola difficilmente calcolabile, mentre la distribuzione B conduce ad espressioni facilmente calcolabili in pratica, come appare dal lavoro di C. Gini e G. Livada (3).

---

(1) C. Gini, *Corso di statistica*, a cura di S. Gatti e C. Benedetti, 1954-55 nuova edizione aggiornata, Veschi, Roma.

(2) C. Gini, *Considerazioni sulle probabilità a posteriori ed applicazioni al rapporto dei sessi nelle nascite umane*, Cagliari 1911 (Riprodotto in « Metron » 1950).

(3) C. Gini e G. Livada, *Sulla probabilità inversa nel caso di grandezze intensive ed in particolare sulle sue applicazioni a collaudi per masse a mezzo di campioni*, « Atti della VI e VII Riun. Scientif. della Soc. Ital. di Statistica », Roma 1943.

L'ipotesi del Gini sulla distribuzione delle probabilità a priori secondo la distribuzione :

$$f_B(x) = \frac{1}{B(l, m)} x^{l-1} (1 - x)^{m-1}; \quad l, m > 1; 0 \leq x \leq 1$$

quindi oltremodo calzante in quanto, come abbiamo fatto notare nei fogli precedenti, si può determinare l ed m in modo da rappresentare in modo assai soddisfacente una legge binomiale quando $n > \frac{1+p}{p}$, $n > \frac{1+q}{q}$.

Diamo, qui di seguito, tre distribuzioni binomiali e le corrispondenti distribuzioni beta ottenute applicando i criteri esposti. Calcolati gli indici di accostamento :

$$I = \frac{\Sigma | f_b - f_B |}{n}$$

tra ogni distribuzione binomiale e la corrispondente beta si sono ottenuti i seguenti tre valori :

0,000796 ; 0,000206 ; 0,000066

Le applicazioni fatte anche esaminando il grafico allegato si possono ritenere quindi soddisfacenti.

SUMMARY

In this work the author shows some simple methods for representing with a satisfactory approximation a binomial distribution through a *B*-distribution. In connection with this the author exhibits some considerations of prof. Gini regarding the distribution of the *a priori* probability in Bayes' formula.

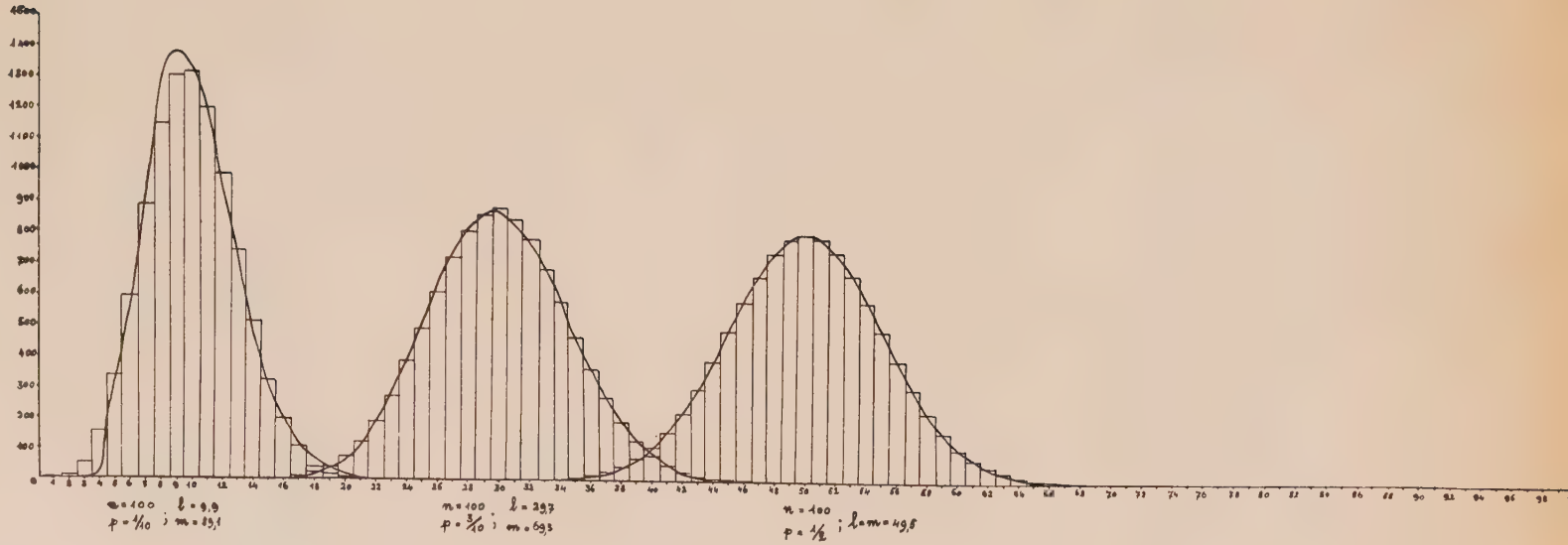
In the above considerations, which give justification to this work, it is observed that, lacking whatsoever knowledge of the *a priori* probability distribution and being forced to take up a decision, the unique hypotesis to be attended to would be that not all the *combinations* were equiprobable but all the *permutations* of the events favorable or contrary to every composition of certain population, because these *permutations* are all equally conceivable ; this does not lead to an equidistribution of the *combinations* and therefore of the *a priori* probability but to an equidistribution of the *permutations* therefore to a binomial distribution.

The *B*-distribution is chosen to represent the binomial one in order to make the Bayes' formula more easily handled.

	$10^4 \binom{n}{nx} \cdot p^{nx} q^{n(1-x)}$	$10^4 \frac{1/n}{B(l,m)} x^{l-1} (1-x)^{m-1}$	$10^4 \binom{n}{nx} \cdot p^{nx} q^{n(1-x)}$	$10^4 \frac{1/n}{B(l,m)} x^{l-1} (1-x)^{m-1}$	$10^4 \binom{n}{nx} \cdot p^{nx} q^{n(1-x)}$	$10^4 \frac{1/n}{B(l,m)} x^{l-1} (1-x)^{m-1}$
$n\,x$	$n = 100$ $p = 1/10$	$l = 9{,}9\ ;\ m = 89{,}1$	$n = 100$ $p = 3/10$	$l = 29{,}7\ ;\ m = 69{,}3$	$n = 100$ $p = 1/2$	$l = m = 49{,}5$
0	—	—	—	—	—	—
1	3	—	—	—	—	—
2	16	—	—	—	—	—
3	59	2	—	—	—	—
4	159	11	—	—	—	—
5	339	323	—	—	—	—
6	596	571	—	—	—	—
7	889	999	—	—	—	—
8	1148	1265	—	—	—	—
9	1304	1378	—	—	—	—
10	1319	1328	—	—	—	—
11	1199	1161	—	—	—	—
12	988	929	—	—	—	—
13	743	767	—	—	—	—
14	513	484	1	—	—	—
15	327	319	2	1	—	—
16	193	200	5	3	—	—
17	106	120	11	8	—	—
18	54	68	23	18	—	—
19	26	37	43	37	—	—
20	12	20	75	70	—	—
21	5	10	123	120	—	—
22	2	5	190	190	—	—
23	1	2	276	283	—	—
24	—	—	389	392	—	—
25	—	—	495	512	—	—
26	—	—	612	631	—	—
27	—	—	720	737	—	—
28	—	—	804	816	—	—

nx	$10^4 \binom{n}{nx} \cdot p^{nx} q^{n(1-x)}$	$10^4 \frac{1/n}{B(l,m)} x^{l-1} (1-x)^{m-1}$	$10^4 \binom{n}{nx} \cdot p^{nx} q^{n(1-x)}$	$10^4 \frac{1/n}{B(l,m)} x^{l-1} (1-x)^{m-1}$	$10^4 \binom{n}{nx} \cdot p^{nx} q^{n(1-x)}$	$10^4 \frac{1/n}{B(l,m)} x^{l-1} (1-x)^{m-1}$
	$n = 100$ $p = 1/10$	$l = 9{,}9$; $m = 89{,}1$	$n = 100$ $p = 3/10$	$l = 29{,}7$; $m = 69{,}3$	$n = 100$ $p = 1/2$	$l = m = 49{,}5$
29	—	—	855	860	—	—
30	—	—	868	863	—	—
31	—	—	840	828	—	—
32	—	—	776	760	1	1
33	—	—	685	678	2	2
34	—	—	579	563	4	3
35	—	—	467	456	9	9
36	—	—	362	355	16	15
37	—	—	268	266	27	26
38	—	—	191	191	45	45
39	—	—	132	133	71	71
40	—	—	84	89	108	109
41	—	—	53	56	158	161
42	—	—	32	36	222	226
43	—	—	18	21	299	303
44	—	—	10	12	388	392
45	—	—	5	7	484	486
46	—	—	2	4	579	579
47	—	—	1	2	666	665
48	—	—	—	1	737	732
49	—	—	—	—	782	779
50	—	—	—	—	798	791
51	—	—	—	—	782	779
52	—	—	—	—	737	732
53	—	—	—	—	666	665
54	—	—	—	—	579	579
55	—	—	—	—	484	486
56	—	—	—	—	388	392
57	—	—	—	—	299	303

$n\,x$	$10^4\binom{n}{nx}\cdot p^{nx}q^{n(1-x)}$	$10^4\,\frac{1/n}{B(l,m)}\,x^{l-1}\,(1-x)^{m-1}$	$10^4\binom{n}{nx}\cdot p^{nx}q^{n(1-x)}$	$10^4\,\frac{1/n}{B(l,m)}\,x^{l-1}\,(1-x)^{m-1}$	$10^4\binom{n}{nx}\cdot p^{nx}q^{n(1-x)}$	$10^4\,\frac{1/n}{B(l,m)}\,x^{l-1}\,(1-x)^{m-1}$
	$n = 100$ $p = 1/10$	$l = 9{,}9\ ;\ m = 89{,}1$	$n = 100$ $p = 3/10$	$l = 29{,}7\ ;\ m = 69{,}3$	$n = 100$ $p = 1/2$	$l = m = 49{,}5$
58	—	—	—	—	222	226
59	—	—	—	—	158	161
60	—	—	—	—	108	109
61	—	—	—	—	71	71
62	—	—	—	—	45	45
63	—	—	—	—	27	26
64	—	—	—	—	16	15
65	—	—	—	—	9	9
66	—	—	—	—	4	3
67	—	—	—	—	2	2
68	—	—	—	—	1	1
69	—	—	—	—	—	—
70	—	—	—	—	—	—
71	—	—	—	—	—	—
72	—	—	—	—	—	—
73	—	—	—	—	—	—
74	—	—	—	—	—	—
75	—	—	—	—	—	—
76	—	—	—	—	—	—
77	—	—	—	—	—	—
78	—	—	—	—	—	—
79	—	—	—	—	—	—
80	—	—	—	—	—	—
81	—	—	—	—	—	—
82	—	—	—	—	—	—
83	—	—	—	—	—	—
84	—	—	—	—	—	—
85	—	—	—	—	—	—
.	—	—	—	—	—	—
.	—	—	—	—	—	—
.	—	—	—	—	—	—
.	—	—	—	—	—	—
100	—	—	—	—	—	—


n = 100
p = 1/10 ; l = 9,9 ; m = 89,1
n = 100
p = 3/10 ; l = 29,7 ; m = 69,3
n = 100
p = 1/2 ; l = m = 49,5

TOMMASO SALVEMINI

Varianza della differenza media dei campioni ottenuti secondo lo schema di estrazione in blocco (*)

1. — Premessa.

Gli errori quadratici medi degli indici statistici (o costanti caratteristiche) sono calcolati, generalmente, nell'ipotesi di campioni ricavati secondo lo schema delle prove ripetute, cioè mediante *estrazione a caso con ripetizione.*

Ora è da osservare che, nella pratica statistica, quando da una popolazione di H elementi si estrae a caso un campione di N elementi, si fa in realtà una scelta *senza ripetizione.*

Questo schema probabilistico è detto dell'*estrazione in blocco,* essendo indifferente estrarre gli elementi uno alla volta o tutti insieme. Esso è più complicato dell'altro perchè le singole prove non sono indipendenti, in quanto cambia di volta in volta la composizione dell'urna in base ai risultati ottenuti nelle prove precedenti.

Forse in relazione a tale complessità, o per i motivi che dirò tra poco, soltanto per alcune costanti statistiche si conosce l'errore quadratico medio nell'ipotesi di estrazione in blocco (1).

(*) Comunicazione presentata al Seminario di Statistica Metodologica della Facoltà di Scienze Statistiche, Dem. ed Attuariali della Università di Roma il 3 Marzo 1956.

(1) La formula della varianza della media per campioni ottenuti con detto schema trovasi in G. Mortara, *Elementi di Statistica,* Athenaeum, 1917, a pag. 356 ; una trattazione ampia in cui si ricavano i valori medi dei momenti inerenti a detto schema è quella di Al.A. Tschuprow, *On the mathematical expectation of the moments of frequency distributions in the case of correlated observations,* « Metron », Vol. II, 1923. Vedasi pure : J.S. Neyman, *Contributions to the theory of small samples drawn from a finite population,* « Biometrika », XVII, 1925 ; A.E.R. Church, *On the means and*

Per le altre si ricorre alla comoda proprietà che i risultati dello schema dell'estrazione in blocco differiscono poco da quelli dell'estrazione con ripetizione quando la massa dei casi, da cui il campione viene estratto, tende ad essere infinitamente numerosa, oppure quando il campione contiene una piccola frazione della massa finita di elementi, perchè, in tal caso, si ritiene che il trascurare un numero relativamente piccolo di questi elementi non abbia effetto sensibile sulla costituzione della restante popolazione (2).

Questo ragionamento ha, evidentemente, bisogno di essere precisato matematicamente. A tale scopo occorre determinare la relazione tra H, N e l'errore che si commette sostituendo, nel procedimento di calcolo, lo schema di Bernoulli a quello dell'estrazione in blocco. In altri termini ci si può chiedere: quanto piccolo deve essere il rapporto $\frac{N}{H}$ affinchè l'errore quadratico medio di un qualunque indice statistico calcolato seguendo lo schema della estrazione in blocco possa essere sostituito, con assegnata approssimazione, col risultato più semplice proveniente dallo schema bernoulliano? Ora è chiaro che dare una relazione a carattere generale, valida, cioè, per tutti gli indici, è difficile, mentre può essere utile vedere se le conclusioni buone per alcuni casi particolari, sia pure importanti, siano valide anche per altri casi.

D'altra parte, le espressioni ricavate con lo schema di estrazione a caso senza ripetizione o in blocco, avendo come limite, per $N \to \infty$, quelle dell'altro schema usuale dell'estrazione con ripetizione, sono più generali di queste. La loro importanza teorica e pratica è quindi indubbia.

---

squared standard deviations of small samples from any population, « Biometrika », XVIII, 1926; C. Gini, *Le medie dei campioni*, « Metron », XV, 1950; V. Castellano, *Introduzione alla teoria dei campioni*, « Statistica », XI, 3-4, 1951; G. Pompilj, *Complementi di calcolo delle probabilità*, Veschi, Roma, 1948; *Sulle medie combinatorie potenziate dei campioni*, « Rendiconti Seminario matematico di Padova », XVIII, 1949; *Sulla media e la varianza di un campione*, « Gazeta de matematica », Lisbona, Dez. 1951, N° 50.

(2) M.G. Kendall, *The advanced theory of statistics*, Griffin, London, 1948, Vol. 1, pag. 186.

In questa nota mi limiterò a ricavare l'espressione della varianza della differenza media nello schema detto e a farne un confronto con quella ricavata in base allo schema bernoulliano (3). Da tale confronto si potrà ricavare una risposta rigorosa alla precedente domanda; tuttavia l'espressione piuttosto complessa in base alla quale può ricavarsi la risposta ci indurrà a sostituire l'espressione esatta con altra approssimata che ci sembra sufficiente allo scopo.

Un'applicazione concreta mostrerà l'influenza che ha sul risultato la sostituzione delle quantità riguardanti la massa con stime ricavate dai campioni.

2. — La differenza semplice media della massa e del campione.

Sia data una massa finita di H elementi (non necessariamente tutti distinti)

$$x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_H \qquad (1)$$

---

(3) L'errore quadratico medio della differenza media è stato calcolato per la prima volta — come nota C. Gini nelle *Memorie di metodologia statistica* (Giuffré, Milano, 1939, pag. 250) — da F.R. Helmert nell'ipotesi della distribuzione normale (*Die Genauigkeit der Formel von Peters zur Berechnung des Wahrscheinlichen Beobachtungs-fehlers directer Beobachtungen gleicher Genauigkeit*, « Astronomische Nachrichten », n. 2096, 1876). Successivamente, per distribuzioni statistiche qualunque, e sempre nell'ipotesi di estrazioni con ripetizione, furono date espressioni complicate e, a volte, approssimate da U.S. Nair (*The standard error of Gini's mean difference*, « Biometrika », 28, 1936, pag. 428), A.L. Bowley (*Standard Deviation of Gini's mean difference*, « Comptes rendus du congres international des mathematiciens », Oslo, 1936, pagg. 182-83), H. Wold, (*On the mean difference at random samples. A note on prof. Bowley's lecture*, Ibidem, pagg. 212-13), Z.A. Lomnicki, (*The standard error of Gini's mean difference*, « Annals of Math. statistics », Vol. 23, 4, 1952), mentre una espressione esatta, più rispondente per le applicazioni, trovasi in B. Michetti, (*Metodi vari per la misura della variabilità dei fenomeni*, Tesi di laurea della Facoltà di Econ. e Comm. dell'Università di Roma, 1948). Richiameremo nel seguito quest'ultimo risultato e quello di Lomnicki.

e sia

$$\Delta = \frac{2}{H\,(H-1)} \sum_{i>j} (x_i - x_j) \qquad (2)$$

la differenza semplice media (senza ripetizione) (4).

Supponiamo ora di prelevare N degli H elementi secondo lo schema di estrazione in blocco. Ogni x_i *non* potrà figurare più di una volta tra gli N prelevati.

È evidente che a ciascuna di queste quantità possiamo associare una v.c. (sigla di variabile casuale) elementare $\mathbf{O}_i$ che assume i valori 1 e 0 rispettivamente con probabilità $\frac{N}{H}$ e $\frac{H-N}{H}$; indichiamo con o_i una sua determinazione particolare. (5) Potremo scrivere, per la differenza semplice media nel campione,

$$_c\Delta = \frac{2}{N\,(N-1)} \sum_{ij} (x_i - x_j)\, o_i\, o_j\,. \qquad (3)$$

È ora noto che i campioni distinti ricavabili secondo lo schema detto sono $\binom{H}{N}$, ciascuno con eguale probabilità di essere estratto, i quali costituiscono l'universo dei campioni. Pertanto, in analogia a quanto fa il Pompilj in altro caso (6), alle quantità $_c\Delta$ possiamo associare una v.c. $\mathbf{D}$ espressa da

$$\mathbf{D} = \frac{2}{N\,(N-1)} \sum_{i>j} (x_i - x_j)\, \mathbf{O}_i\, \mathbf{O}_j\,. \qquad (4)$$

---

(4) C. Gini, *Variabilità e mutabilità*, Studi economico-giuridici della Università di Cagliari, 1912; ripubblicata in *Memorie di metodologia statistica*, Vol. I, *Variabilità e concentrazione*, 1ª edizione, Giuffrè, Milano, 1939, 2ª edizione con note di aggiornamento a cura di E. Pizzetti e T. Salvemini, Veschi, Roma, 1955.

(5) Le v.c. sono indicate con carattere neretto, mentre le determinazioni particolari sono in carattere corsivo.

(6) G. Pompilj, *Sulla media geometrica e sopra un indice di mutabilità calcolati mediante un campione*, «Memorie della Soc. It. delle Scienze detta dei XL», serie IIIª, vol. XXVI, 1947; *Complementi di calcolo delle probabilità*, Veschi, Roma, anno accademico, 1948-49.

È stato già dimostrato da C. GINI (7) in base a ragionamento diretto sul modo di estrazione del campione, che il valor medio della v.c. **D** è:

$$M(\mathbf{D}) = \Delta \qquad (5)$$

Questa semplice relazione ha importanza per quel che segue e può avere altra immediata verifica osservando che, dalla (4),

$$M(\mathbf{D}) = \frac{2}{N(N-1)} \sum_{i>j} (x_i - x_j)\, M(\mathbf{O}_i\, \mathbf{O}_j)$$

e tenendo presente, come richiameremo tra breve [formula (b) della nota (11)] che

$$M(\mathbf{O}_i\, \mathbf{O}_j) = \frac{N(N-1)}{H(H-1)},$$

per cui, sostituendo e semplificando, si ha la (5).

Prendendo, quindi, la differenza semplice media senza ripetizione tanto nella massa quanto nel campione, il valore medio della v.c. **D** coincide col valore che ha detto indice nella massa da cui il campione proviene.

Non altrettanto avviene se si prende la differenza semplice media con ripetizione Δ_R, di cui indichiamo con $\mathbf{D}_R$ la v.c. costituita dalle varie stime, giacchè (GINI, op. cit. nota prec.),

$$M(\mathbf{D}_R) = \frac{N-1}{N} \frac{H}{H-1} \Delta_R .$$

Le stime di Δ_R sono quindi affette da un errore sistematico (8).

Lo stesso si può dire, sia pure con diversa intensità del fattore di correzione, quando i campioni siano ottenuti mediante estrazioni secondo lo schema bernoulliano e si prenda la differenza semplice media, senza o con ripetizione, tanto per il campione che per la massa. Si dimostra, però, che nell'ultima

---

(7) C. GINI, *Variabilità e mutabilità*, op. cit., pag. 256 della edizione 1955.

(8) Gli anglosassoni usano il termine «*bias*» che letteralmente può essere tradotto con *tendenziosità* o *distorsione*, ma questi vocaboli ci sembrano meno appropriati del termine *errore sistematico* (giacchè può essere eliminato, in media, con una correzione sulle singole stime).

ipotesi ammessa per la formazione dei campioni, la differenza semplice media *senza ripetizione* è priva di errore sistematico se nella massa si prende la differenza media *con ripetizione* (9).

3. — La varianza di **D**.

Dobbiamo ora approfondire lo studio della v.c. **D** descritta da ${}_c\Delta$ al variare del campione, cioè lo studio delle stime ${}_c\Delta$. Molto complicata si presenta la ricerca della funzione di ripartizione di tale v.c.; tale problema non è stato finora risolto nemmeno per campioni ottenuti secondo lo schema bernoulliano (10). Dobbiamo quindi contentarci di conoscere, oltre la media, la varianza per avere il grado di attendibilità dell'indice calcolato nel campione (cioè, per avere un numero che permetta di vedere in quale misura il valore ottenuto per il campione si possa scostare dal suo valore medio). Questo è, come si sa, *un problema diretto* di statistica pura, perchè si tratta di calcolare la varianza di **D** inerente all'insieme di tutti i campioni che si possono ricavare dalla massa data secondo lo schema dell'estrazione in blocco.

Indichiamo con $\mathbf{S}_D$ la v.c. scarto tra la **D** e il suo valor medio. Poniamo, cioè,

$$\mathbf{S}_D = \mathbf{D} - \Delta = \tag{6}$$

$$= \frac{2}{N(N-1)} \sum_{i>j} (x_i - x_j)\, \mathbf{O}_i\, \mathbf{O}_j - \frac{2}{H(H-1)} \sum_{i>j} (x_i - x_j)\,.$$

---

(9) Infatti

$$M\left[\frac{2}{N(N-1)} \sum_{i>j} (x_i - x_j)\, \mathbf{A}_i\, \mathbf{A}_j\right] =$$

$$= \frac{2}{N(N-1)} \sum_{i>j} (x_i - x_j)\, M\,(N p_i + \mathbf{L}_i)\,(N p_j + \mathbf{L}_j) = 2 \sum_{i>j} (x_i - x_j)\, p_i\, p_j$$

dove $\mathbf{A}_i$ è la v.c. descritta dalla frequenza a_i di x_i al variare del campione, $\mathbf{L}_i = \mathbf{A}_i - N p_i$ è la v.c. scarto ed è (cfr. G. Pompilj, *Complementi di calcolo delle probabilità*, 1948, p. 124) $M(\mathbf{L}_i) = o$, $M(\mathbf{L}_i\, \mathbf{L}_j) = -N p_i p_j$, dove p_i e p_j sono le probabilità di x_i e x_j nella v.c. divisa in intervalli associata alla massa.

(10) Nel caso di distribuzione normale si è giunti al calcolo del momento terzo. Vedasi: A.R. Kamat, *The third moment of Gini's mean difference*, «Biometrika», Vol. 40, Dec. 1953.

Per quanto abbiamo già visto è chiaro che

$$M(\mathbf{S}_D) = 0 . \tag{7}$$

Ciò posto, indichiamo con

$$\mathbf{R}_i = \mathbf{O}_i - \frac{N}{H} \tag{8}$$

la generica v.c. scarto che assume i valori $\frac{H-N}{H}$ e $\frac{-N}{H}$ quando $\mathbf{O}_i$ assume rispettivamente i valori 1 e 0.

Posto $\alpha = \frac{N}{H}$, $\beta = 1 - \alpha$, si hanno le seguenti formule che saranno utili nel seguito [11]:

$$\left\{\begin{aligned}
&M(\mathbf{R}_i^2) = \alpha\beta \\
&M(\mathbf{R}_i \mathbf{R}_j) = -\frac{\alpha\beta}{H-1} \\
&M(\mathbf{R}_i^3) = \alpha\beta(\beta-\alpha) \\
&M(\mathbf{R}_i^2 \mathbf{R}_j) = -\frac{\alpha\beta(\beta-\alpha)}{H-1} \\
&M(\mathbf{R}_i \mathbf{R}_j \mathbf{R}_w) = 2\frac{\alpha\beta(\beta-\alpha)}{(H-1)(H-2)} \\
&M(\mathbf{R}_i^4) = \alpha\beta(1-3\alpha\beta) \\
&M(\mathbf{R}_i^3 \mathbf{R}_j) = -\frac{\alpha\beta(1-3\alpha\beta)}{H-1} \\
&M(\mathbf{R}_i^2 \mathbf{R}_j^2) = \alpha\beta\left[\alpha\beta - \frac{(\beta-\alpha)^2}{H-1}\right] \\
&M(\mathbf{R}_i^2 R_j \mathbf{R}_w) = \frac{\alpha\beta}{H-1}\left]2\frac{(\beta-\alpha)^2}{H-2} - \alpha\beta\right] \\
&M(\mathbf{R}_i \mathbf{R}_j \mathbf{R}_w \mathbf{R}_z) = 3\frac{\alpha\beta\{(N-1)\beta - (\beta-\alpha)(\beta-2\alpha)\}}{(H-1)(H-2)(H-3)} .
\end{aligned}\right. \tag{9}$$

---

[11] G. Pompilj, *Sulla media e la varianza di un campione*, « Gazeta de matematica », N. 50, dic. 1951. Alle formule (9) si arriva col seguente ragionamento : per calcolare la probabilità che la v.c. *k-pla* ($k \leq N$) ($\mathbf{O}_{i1}, \mathbf{O}_{i2}, \ldots \ldots \mathbf{O}_{ik}$), legata al numero di volte in cui nelle N prove si presentano k elementi determinati, assuma la *k-pla* $(1, 1, \ldots, 1)$ — cioè che tutti i k elementi

Sostituendo nella (6) ad $\mathbf{O}_i$ e $\mathbf{O}_j$ i valori che si ricavano dalla (9) potremo scrivere $\mathbf{S}_D$ in funzione delle v.c. scarto $\mathbf{R}_i$ ed $\mathbf{R}_j$:

$$\mathbf{S}_D = \frac{H-N}{H(N-1)}\Delta + \frac{2}{H(N-1)}\left[\sum_{i>j}(x_i - x_j)\,\mathbf{R}_i + \sum_{i>j}(x_i - x_j)\,\mathbf{R}_j\right] + \tag{10}$$

$$+ \frac{2}{N(N-1)}\sum_{i>j}(x_i - x_j)\,\mathbf{R}\,\mathbf{R}_{ij}$$

da cui

$$\mathbf{S}_D^2 = \mathbf{S}_D\,\mathbf{S}_D = \frac{H-N}{H(N-1)}\Delta\,\mathbf{S}_D + \frac{2}{H(N-1)}\left[\sum_{i>j}(x_i - x_j)\,\mathbf{R}_i\,\mathbf{S}_D + \right. \tag{11}$$

$$\left. + \sum_{i>j}(x_i - x_j)\,\mathbf{R}_j\,\mathbf{S}_D\right] + \frac{2}{N(N-1)}\sum_{i>j}(x_i - x_j)\,\mathbf{R}_i\,\mathbf{R}_j\,\mathbf{S}_D$$

Il calcolo del valor medio di questa quantità è quindi ricondotto a quello di $M(\mathbf{S}_D)$, $M(\mathbf{R}_i\,\mathbf{S}_D)$ o $M(\mathbf{R}_j\,\mathbf{S}_D)$ e $M(\mathbf{R}_i\,\mathbf{R}_j\,\mathbf{S}_D)$ dei quali già sappiamo che $M(\mathbf{S}_D) = 0$. Per il calcolo degli altri valori medi partiamo ancora dalla (10) moltiplicata per una $\mathbf{R}_t$

---

si trovino tra gli N prelevati — si osservi che i gruppi di N elementi che si possono formare con gli H dati sono pari alle disposizioni di H elementi della classe N, $D_{H,N}$; tra questi, quei gruppi che contengono i k elementi considerati, disposti a k a k in tutti i modi possibili tra gli N prelevati, sono pari a tante volte le disposizioni di $H-k$ elementi di classe $N-k$ quante sono le disposizioni di N elementi di classe k, cioè $D_{N,k}\cdot D_{H-k,N-k}$; pertanto, la probabilità cercata è

(a) $$\frac{D_{N,k}\cdot D_{H-k,N-k}}{D_{H,N}} = \frac{N(N-1)\ldots(N-k+1)}{H(H-1)\ldots(H-k+1)}$$

Per $k = 1, 2, 3, 4$, si ottiene

(b) $$M(\mathbf{O}_i^r) = \frac{N}{H},\quad M(\mathbf{O}_i^r\,\mathbf{O}_j^s) = \frac{N(N-1)}{H(H-1)},\quad M(\mathbf{O}_i^r\,\mathbf{O}_j^s\,\mathbf{O}_u^t) = \frac{N(N-1)(N-2)}{H(H-1)(H-2)}$$

$$M(\mathbf{O}_i^r\,\mathbf{O}_j^s\,\mathbf{O}_u^t\,\mathbf{O}_z^w) = \frac{N(N-1)(N-2)(N-3)}{H(H-1)(H-2)(H-3)}$$

quali che siano u, s, t, w, purchè non negativi.

Ciò posto, le formule (9) si ricavano facilmente ove si sostituisca in esse $\mathbf{R}_i$ con $\mathbf{O}_i - \frac{N}{H}$, si sviluppi la potenza o il prodotto corrispondente e si tenga conto dei precedenti valori medi.

generica, avendo però cura di distinguere, come è facile vedere dallo sviluppo successivo, il caso in cui $\mathbf{R}_t$ coincide con $\mathbf{R}_i$ da quello in cui coincide con $\mathbf{R}_j$. Si ha:

$$M(\mathbf{S}_D \mathbf{R}_t) = \frac{H-N}{H(N-1)} \Delta \cdot M(\mathbf{R}_t) + \frac{2}{H(N-1)} \left\{ \sum_{i>j}^{i \neq t} (x_i - x_j) \cdot M(\mathbf{R}_i \mathbf{R}_t) + \right.$$

$$+ \sum_{j}^{i>j} (x_t - x_j) \cdot M(\mathbf{R}_t^2) + \sum_{i>j}^{j \neq t} (x_i - x_j) M(\mathbf{R}_j \mathbf{R}_t) + \sum_{i}^{i>t} (x_i - x_t) \cdot M(\mathbf{R}_t^2) \Big\} +$$

$$+ \frac{2}{N(N-1)} \left\{ \sum_{i>j}^{i,j \neq t} (x_i - x_j) \cdot M(\mathbf{R}_i \mathbf{R}_j \mathbf{R}_t) + \sum_{i}^{i>t} (x_i - x_t) M(\mathbf{R}_i \mathbf{R}_t^2) + \right.$$

$$\left. + \sum_{j}^{t>j} (x_t - x_j) \cdot M(\mathbf{R}_t^2 \mathbf{R}_j) \right\}$$

nelle cui sommatorie t è costante ed i e j variano da 1 ad H, soddisfacendo però le limitazioni indicate nella parte superiore delle sommatorie. Valendoci ora dei risultati (9) si ha, dopo varie semplificazioni,

$$M(\mathbf{S}_D \mathbf{R}_t) = \frac{-2(H-N)}{H(H-2)} \Delta + \tag{12}$$

$$+ \frac{2(H-N)}{H(H-1)(H-2)} \left[\sum_{i}^{i>t} (x_i - x_t) + \sum_{j}^{t>j} (x_t - x_j) \right].$$

In modo analogo si procede per ricavare $M(\mathbf{S}_D \mathbf{R}_t \mathbf{R}_s)$, pur essendo i calcoli più complessi dovendo tener distinti i casi di diseguaglianza tra gli indici i, j, t, ed s da quelli in cui t ed s coincidono con i o con j separatamente o simultaneamente, per cui, in conclusione, si dovranno separare 13 termini di cui 11 contengono sommatorie. Applicando poi le (9) e semplificando opportunamente si ricava (omettendo per brevità i passaggi):

$$M(\mathbf{S}_D \mathbf{R}_t \mathbf{R}_s) = \frac{(H-N)(6H^2 + 12N - 10HN - 6H)}{H^2(H-1)(H-2)(H-3)} \Delta +$$

$$+ \frac{2(H-N)(3N-2H)}{H^2(H-1)(H-2)(H-3)} \left[\sum_{i}^{i>t} (x_i - x_t) + \sum_{i}^{i \neq t} (x_i - x_s) + \right. \tag{13}$$

$$\left. + \sum_{j}^{i \neq s} (x_t - x_j) + \sum_{j}^{s>j} (x_s - x_j) \right] + \frac{2(H-N)(H-N-1)}{H(H-1)(H-2)(H-3)} (x_t - x_s).$$

Riprendiamo ora la (11) e calcoliamone il suo valor medio che indicheremo — come è ormai nell'uso — col simbolo *Var* (**D**). Tenendo presente la (7), la (12) per $t = i$ e $t = j$, e la (13) si ha, dopo qualche semplificazione,

$$Var\,(\mathbf{D}) = M\,(\mathbf{S}_{\mathbf{D}}^2) = \frac{2\,(H-N)\,(3\,H+3\,N-3-2\,H\,N)}{N\,(N-1)\,(H-2)\,(H-3)}\,\Delta^2 +$$

$$+\frac{4\,(H-N)\,(N-2)}{N\,(N-1)\,H\,(H-1)\,(H-2)\,(H-3)}\Big[\sum_{i=1}^{H}\sum_{j=1}^{i}(x_i-x_j)\sum_{l=1}^{i}(x_i-x_l)+ \tag{14}$$

$$+\sum_{j=1}^{H}\sum_{i=j+1}^{H}(x_i-x_j)\sum_{l=j+1}^{H}(x_l-x_j)+2\sum_{j=1}^{H}\sum_{i=j+1}^{H}(x_i-x_j)\sum_{l=1}^{j}(x_j-x_l)\Big]+$$

$$+\frac{4\,H\,(H-N)\,(H-N-1)}{N\,(N-1)\,(H-1)\,(H-2)\,(H-3)}\,\sigma^2\,.$$

Per semplificare questa espressione, indichiamo rispettivamente con A, B e C le tre somme triple che figurano nella parentesi quadra. Queste quantità (come del resto anche la varianza e Δ) sono invarianti per traslazione dell'origine, per cui senza limitare la generalità dei risultati, possiamo supporre che l'origine degli assi sia la media aritmetica, così che

$$\sum_{h=i+1}^{H} x_h = -\sum_{h=1}^{i} x_h$$

Sviluppando allora due delle sommatorie che figurano in A, B e C e valendoci della precedente relazione, dopo qualche semplice trasformazione, potremo scrivere:

$$\left.\begin{aligned}
A &= \sum_{i=1}^{H}\Big[i^2 x_i^2 - 2\,i\,x_i\sum_{h=1}^{i}x_h+\Big(\sum_{h=1}^{i}x_h\Big)^2\Big]\\
B &= \sum_{i=1}^{H}\Big[i^2 x_i^2+2\,(H-i)\,x_i\sum_{h=1}^{i}x_h+\Big(\sum_{h=1}^{i}x_h\Big)^2-2\,H\,i\,x_i^2+H^2x_i^2\Big]\\
C &= \sum_{i=1}^{H}\Big[i^2 x_i^2 - 2\,i\,x_i\sum_{h=1}^{i}x_h+\Big(\sum_{h=1}^{i}x_h\Big)^2-H\,i\,x_i^2+H\,x_i\sum_{1}^{i}x_h\Big],
\end{aligned}\right\} \tag{15}$$

dalle quali si vede immediatamente che la quantità tra parentesi quadra della (14) si esprime in funzione di C e di σ^2 nel seguente modo:

$$A + B + 2C = 4C + H^3\sigma^2 . \tag{16}$$

Pertanto

$$Var(\mathbf{D}) = \frac{2(H-N)}{N(N-1)(H-1)(H-2)(H-3)} \cdot$$
$$\cdot \left[(H-1)(3H + 3N - 3 - 2HN)\Delta^2 + \right. \tag{17}$$
$$\left. + 2H^2(N-2)\left(\frac{4C}{H^3} + \sigma^2\right) + 2H(H-N-1)\sigma^2 \right]$$

che, con opportuna trasformazione, possiamo mettere, in definitiva, sotto la seguente forma, da cui è più facile fare deduzioni:

$$Var(\mathbf{D}) = \frac{2}{N}\frac{H-N}{H-3}\left[-\frac{2H-3}{H-2}\Delta^2 + \frac{4}{3}\Delta_{21} + \frac{2H^2}{(H-1)(H-2)}\sigma^2 \right] +$$
$$+ \frac{2}{N(N-1)}\frac{H-N}{H-3}\left[\frac{H}{H-2}\Delta^2 - \frac{4}{3}\Delta_{21} - \frac{2H(N+1)}{(H-1)(H-2)}\sigma^2 \right] \tag{18}$$

dove

$$\Delta_{21} = \frac{6}{H(H-1)(H-2)}C = \frac{6}{H(H-1)(H-2)}\sum_{j=1}^{H}\sum_{i=j+1}^{H}(x_i - x_j) \cdot$$
$$\cdot \sum_{l=1}^{j}(x_j - x_l) = \frac{6}{H(H-1)(H-2)}\sum_{i=1}^{H}\left(i\,x_i - \sum_{l=1}^{i} x_l \right)\left(\sum_{l=i+1}^{H} x_l - (H-i)\,x_i \right), \tag{19}$$

Questo Δ_{21} può ottenersi anche sostituendo a C l'espressione che questa quantità ha nella (15) quando le x_i hanno, come in quel caso, medio zero. Δ_{21} può essere considerato un indice di variabilità, ed essere chiamato *differenza media di ordine* 2, perchè, in sostanza, contiene sotto sommatorie il prodotto di due differenze anzicchè una sola come nella differenza semplice media; il suo significato statistico non interessa la presente ricerca e perciò non ci soffermiamo su esso. Determineremo successivamente la media della v.c. descritta dai valori di Δ_{21} ricavati dai campioni.

4. — Considerazioni riguardanti l'espressione di Var (**D**).

Dalla (18) è facile vedere che per N fissò ed H crescente si ha :

$$\lim_{H\to\infty} Var\,(\mathbf{D}) = \frac{2}{N}\left[-2\,\Delta^2 + \frac{4}{3}\,\Delta_{21} + 2\,\sigma^2\right] + \frac{2}{N\,(N-1)}\left[\Delta^2 - \frac{4}{3}\,\Delta_{21}\right]. \tag{18'}$$

Ricordiamo, poi, che indicando con Var (**D**)$_r$ la varianza della v.c. **D** descritta dai valori che la differenza semplice media prende nei campioni ricavati secondo lo schema dell'estrazione con ripetizione, si ha [12]

$$Var\,(\mathbf{D})_r = \frac{2}{N}\left(-2\,\Delta_R^2 + \frac{4}{3}\,\Delta_{21,R} + 2\,\sigma^2\right) + \frac{2}{N\,(N-1)}\left[\Delta_R^2 - \frac{4}{3}\,\Delta_{21,R}\right] \tag{20}$$

dove $\Delta_{21,R}$ è il termine corrispondente di Δ_{21} quando, nella (19) al posto di $H\,(H-1)\,(H-2)$ sostituiamo H^3 (e, per analogia con

---

[12] B. Michetti, op. cit. in nota (3).
Il Lomnicki (1952, op. cit.) dà la seguente espressione :

$$Var\,(\mathbf{D}) = \frac{2}{N\,(N-1)}\left\{2\,(N-1)\,\sigma^2 + 8\,(N-2)\,I - (2\,N-3)\,\Delta_R^2\right\}$$

dove :

$$I = \int_{-\infty}^{+\infty}\int_{-\infty}^{x}\int_{x}^{\infty}(x-y)\,(z-x)\,f\,(x)\,f\,(y)\,f\,(z)\,d\,x\,d\,y\,d\,z\,.$$

Essa coincide, sostanzialmente, con quella del Michetti (1948) : basta, infatti, fare una semplice trasformazione algebrica e tener presente che l'integrale triplo I, riferito ad una v.c. divisa in intervalli, non è altro che $\frac{1}{6}\,\Delta_{21,R}$, potendo scrivere

$$\Delta_{21,R} = 6\sum_{i>j>l}(x_i - x_j)\,(x_j - x_l)\,p_i\,p_j\,p_l$$

in cui p_i, p_j e p_l sono le probabilità rispettive di x_i, x_j e x_l (nella massa).

la denominazione di Δ_R, potrebbe essere chiamato differenza media con ripetizione del secondo ordine). Ora, poichè per H sufficientemente grande Δ_{2i} e $\Delta_{21,R}$, così come Δ e Δ_R, differiscono pochissimo tra loro, potremo scrivere

$$\lim_{H \to \infty} Var\,(\mathbf{D}) = Var\,(\mathbf{D})_r . \qquad (20')$$

Quest'ultima quantità è dunque, come avevamo detto, deducibile da quella più generale da noi data.

Per $N = H$ la (18) si annulla come era da aspettarsi, perchè, ovviamente, il campione coincide in tal caso con l'intera popolazione.

Ricordiamo ora, come abbiamo già accennato nel n. 1, che il risultato valevole nell'ipotesi dello schema bernoulliano si ritiene valido, indipendentemente dalla grandezza di H, quando $\alpha = \frac{N}{H}$ sia sufficientemente piccolo. Questo ha bisogno di essere precisato, calcolando l'approssimazione in relazione al valore di detto rapporto.

Poniamo, nei termini della (18) non compresi nella (20), $N = H\alpha$; si ha

$$Var\,(\mathbf{D}) = \frac{2}{N}(1-\alpha)\frac{H}{H-3}\left[-\frac{H-3/2}{H-2}2\,\Delta^2 + \frac{4}{3}\Delta_{21} + \right.$$

$$\left. + \frac{H}{H-1}\cdot\frac{H}{H-2}2\,\sigma^2\right] + \frac{2}{N(N-1)}(1-\alpha)\frac{H}{H-3}\left[\frac{H}{H-2}\Delta^2 - \right. \qquad (21)$$

$$\left. - \frac{4}{3}\Delta_{21} - \frac{H}{H-1}\frac{H}{H-2}2\left(\alpha + \frac{1}{H}\right)\sigma^2\right]$$

Se H è sufficientemente grande, tale da poter porre $\frac{H}{H-i} = 1$, per $i = 1$, 2 e 3, e tale da ritenere trascurabili le differenze tra Δ e Δ_R, tra Δ_{21} e $\Delta_{21,R}$, dalla precedente si ha

$$Var\,(\mathbf{D}) \simeq (1-\alpha)\,Var\,(\mathbf{D})_r - \frac{4(1-\alpha)}{N(N-1)}\left(\alpha + \frac{1}{H}\right)\sigma^2 \qquad (22)$$

e se anche N è sufficientemente grande potremo scrivere

$$Var\,(\mathbf{D}) \simeq (1-\alpha)\,Var\,(\mathbf{D})_r = \left(1-\frac{N}{H}\right) Var\,(\mathbf{D})_r \qquad (23)$$

dalla quale si ha modo di valutare approssimativamente la differenza di risultato nel calcolo della varianza della differenza media dei campioni secondo che si usi lo schema di estrazione in blocco o lo schema bernoulliano [13].

Se quindi fissiamo che, nel calcolo della varianza della differenza media dei campioni estratti in blocco, non si voglia un errore superiore a $\frac{1}{10^n}$, usando l'espressione valida per lo schema bernoulliano, occorre che la « frazione di campionamento » soddisfi alla seguente relazione:

$$\alpha < \frac{1}{10^n\, Var\,(\mathbf{D})_r} \qquad (24)$$

La (23) e la (24) permettono di regolarsi caso per caso in relazione all'approssimazione che si vuole.

La (23) è l'espressione più semplice per i comuni bisogni riguardanti i campioni sufficientemente grandi. Da essa trae conferma la più stretta approssimazione tra le due varianze a confronto al decrescere di $\alpha = \frac{N}{H}$.

Per maggiore precisione, la relazione fra $Var\,(\mathbf{D})$ e $Var\,(\mathbf{D})_r$ può essere messa anche sotto altra forma. A tale scopo, sostituiamo nella (18), al posto di Δ e di Δ_{21} le corrispondenti espres-

---

[13] La (23) è perfettamente analoga all'espressione che si ricava per la varianza delle medie dei campioni, secondo l'uno e l'altro schema, quando H ed N sono sufficientemente grandi. Infatti, indicando con $\bar{x}$ la media delle x_i, è noto che per campioni estratti in blocco

$$Var\,(\bar{x}) = \frac{\sigma^2}{N}\,\frac{H-N}{H-1} = \frac{\sigma^2}{N}\,(1-\alpha)\,\frac{H}{H-1} \simeq \frac{\sigma^2}{N}\,(1-\alpha) = (1-\alpha)\,Var\,(\bar{x})_r$$

dove $Var\,(\bar{x})_r$ sta ad indicare la varianza delle medie dei campioni ricavati bernoullianamente.

sioni in funzione di Δ_R e $\Delta_{21,R}$. Dopo facili semplificazioni si ha:

$$Var\,(\mathbf{D}) = \frac{H^2\,(H-N)}{(H-1)\,(H-2)\,(H-3)} \cdot$$

$$\cdot \left\{ \frac{2}{N} \left[-2\,\frac{H-3/2}{H-1}\,\Delta_R^2 + \frac{4}{3}\,\Delta_{21,R} + 2\,\sigma^2 \right] + \right. \qquad (18'')$$

$$\left. + \frac{2}{N\,(N-1)} \left[\frac{H}{H-1}\,\Delta_R^2 - \frac{4}{3}\,\Delta_{21,R} - 2\,\frac{N+1}{H}\,\sigma^2 \right] \right\}.$$

Ora si tenga presente che, nei comuni casi concreti, H supera almeno 10, e, quindi, si può ritenere generalmente trascurabile l'errore che si commette sostituendo $\frac{H-3/2}{H-1}$ e $\frac{H}{H-1}$ con l'unità (tanto più che quest'ultima frazione appartiene alla seconda parte della (18'') che è assai meno preponderante della prima perchè moltiplicata per $\frac{2}{N\,(N-1)}$). Posto

$$K = \frac{H^2\,(H-N)}{(H-1)\,(H-2)\,(H-3)}$$

si ha immediatamente

$$Var\,(\mathbf{D}) \cong K \left[Var\,(\mathbf{D})_r - \frac{4}{N\,(N-1)}\,\frac{N+1}{H}\,\sigma^2 \right]. \qquad (22')$$

Questa relazione esprime $Var\,(\mathbf{D})$ in funzione di $Var\,(\mathbf{D})_r$ e di σ con maggiore approssimazione della (22), pur rimanendo approssimata per difetto come quest'ultima.

Risolvendo, poi, rispetto ad N la diseguaglianza

$$\frac{H^2\,(H-N)}{(H-1)\,(H-2)\,(H-3)} \leq 1$$

si ha:

$$N \geq 6 - \frac{11}{H} + \frac{6}{H^2}$$

la quale ci assicura che K è tanto più piccolo dell'unità quanto più N è maggiore di 6 (oppure quando decresce $H\,(\geq N)$ a pari valore di N). Questa circostanza messa insieme al fatto che la (22') contiene un termine positivo in sottrazione, permette di

dedurre che, in generale, lo schema di estrazioni in blocco, in confronto all'altro, determina, anche per la differenza media, meno dispersione dei risultati ottenuti da tutti i possibili campioni.

Osserviamo, inoltre, che, per N sufficientemente grande, si può trascurare tanto nella (18) quanto nella (20) il termine contenente in evidenza $\frac{2}{N(N-1)}$ e dedurne che, in entrambi gli schemi di estrazione, la varianza della differenza media varia, approssimativamente, in ragione inversa della numerosità del campione, e quindi il valore quadratico medio dello scarto tra i valori di Δ calcolati nel campione e quello calcolato nella v.c. da cui il campione proviene, decresce come $\frac{1}{\sqrt{N}}$.

Essendo dunque

$$\lim_{N \to \infty} Var\,(\mathbf{D}) = 0 \qquad (25)$$

e valendo la (5) si è nelle condizioni per asserire ([14]) che **D** è una *stima consistente* (secondo la terminologia di R.A. Fisher), nel senso che, per N tendente all'infinito, essa converge in probabilità al valore Δ.

Può essere utile osservare, infine, che l'espressione esatta (18), pur apparendo più complicata della (20), richiede, in sostanza, quasi lo stesso tempo di calcolo di quest'ultima, perchè la fatica maggiore è nel calcolo di Δ, Δ_{21} e σ. Una volta ottenute queste quantità (il cui calcolo nei due schemi è quasi identico), la maggiore complicazione della (18) è soltanto in alcuni fattori moltiplicativi. Lo stesso si può dire se si adopera la (18′).

5. — Il valore medio di Δ_{21}.

Per approfondire ulteriormente le differenze tra i due sistemi di formazione dei campioni e per l'utilizzazione pratica dei risultati è necessario conoscere, come avevamo già prean-

---

([14]) S.S. Wilks, *Mathematical Statistics*, Princeton, University Press, New Jersey, 1950.

nunciato, il valore medio della v.c. associata alle varie determinazioni di Δ_{21} secondo l'uno o l'altro metodo, per correggere l'espressione di ${}_c\Delta_{21}$ da eventuali errori sistematici.

Poniamo, come si è fatto per Δ nel n. 2,

$$ {}_c\Delta_{21} = \frac{6}{N(N-1)(N-2)} \sum_{j=1}^{H} \sum_{i=j+1}^{H} (x_i - x_j) \sum_{l=1}^{j} (x_j - x_l) o_i o_j o_l \qquad (26) $$

e indichiamo con $\mathbf{D}_{21}$ la v.c. descritta da ${}_c\Delta_{21}$ al variare del gruppo di N prove. Potremo scrivere

$$ \mathbf{D}_{21} = \frac{6}{N\,(N-1)\,(N-2)} \sum_{j=1}^{H} \sum_{i=j+1}^{H} (x_i - x_j) \sum_{l=1}^{j} (x_j - x_l)\, \mathbf{O}_i\, \mathbf{O}_j\, \mathbf{O}_l \qquad (27) $$

Ora, nello schema dell'estrazione in blocco, essendo

$$ M\,(\mathbf{O}_i\, \mathbf{O}_j\, \mathbf{O}_l) = \frac{N\,(N-1)\,(N-2)}{H\,(H-1)\,(H-2)} \qquad (28) $$

si ha subito

$$ M\,(\mathbf{D}_{21}) = \Delta_{21} \qquad (29) $$

Nello schema bernoulliano, ricordiamo che ogni x_i può presentarsi con frequenza assoluta a_i a cui corrisponde, variando il gruppo di prove, una v.c. $\mathbf{A}_i$. Potremo scrivere

$$ {}_c\Delta_{21}^{(r)} = \frac{6}{N\,(N-1)\,(N-2)} \sum_{i>j>l} (x_i - x_j)\,(x_j - x_l)\; a_i\; a_j\; a_l \qquad (30) $$

Indicando allora con $\mathbf{D}_{21}^{(r)}$ la v.c. descritta da ${}_c\Delta_{21}^{(r)}$ e posto $\mathbf{L}_i = \mathbf{A}_i - N\,p_i$, si ha:

$$ M\,(\mathbf{D}_{21}^{(r)}) = \frac{6}{N\,(N-1)\,(N-2)} \sum_{i>j>l} (x_i - x_j)\,(x_j - x_l) \cdot \\ \cdot M\,(N\,p_i + \mathbf{L}_i)\,(N\,p_j + \mathbf{L}_j)\,(N\,p_l + \mathbf{L}_l) \qquad (31) $$

da cui sviluppando i tre ultimi prodotti e sostituendo le formule (6) del lavoro di G. Pompilj (*Sulla media geometrica...* op. cit.) e semplificando si ha

$$ M\,(\mathbf{D}_{21}') = 6 \sum_{i>j>l} (x_i - x_j)\,(x_j - x_l)\; p_i\; p_j\; p_l \qquad (32) $$

che rappresenta l'espressione di $\Delta_{21,R}$ nella massa.

In conclusione, per avere un ${}_c\Delta_{21}$ privo di errore sistematico, occorre che in entrambi gli schemi di formazione dei campioni si prenda per questa quantità l'espressione (30) o la (26) o una trasformata, come ad es. la (19), ove, però, si sostituisca N al posto di H. Questo risultato è analogo a quello del n. 2 riguardante l'adozione di ${}_c\Delta$ come stima priva di errore sistematico in entrambi gli schemi.

Nel caso dello scarto quadratico medio, ricordiamo che dobbiamo prendere

$$ {}_c\sigma^2 = \frac{H-1}{H} \frac{\Sigma (x_i - \bar{x})^2}{N-1} \qquad (33) $$

se i campioni sono ottenuti con lo schema dell'estrazione in blocco, e la stessa quantità senza il fattore $\frac{H-1}{H}$ nel caso dello schema bernoulliano.

Osserviamo infine che, nonostante gli accorgimenti a cui abbiamo fatto ora cenno, quando alle quantità riguardanti la massa dei casi da cui i campioni sono tratti si sostituiscono le corrispondenti quantità stimate (allo scopo di rendere praticamente utilizzabili le espressioni di queste varianze e avere il grado di attendibilità dei risultati riguardanti i campioni) si ottengono, com'è noto, valori che possono differire sensibilmente da quelli esatti. Tuttavia, tali sostituzioni, con le debite riserve, sono comunemente ammesse ([15]).

Nel prossimo paragrafo cercheremo di vedere in concreto l'influenza di detta sostituzione sul valore della varianza della differenza media.

6. — Applicazione ai campioni delle cinquine estratte nel gioco del lotto.

Un esempio di estrazioni a caso senza ripetizione è quello del gioco del lotto in Italia, nel quale, come è noto, si estrag-

---

([15]) Tra i più recenti lavori vedasi: L. Galvani, *Révision critique de certains points de la méthode représentative*, « Revue de l'Inst. intern. de statistique », 1951, 1 e *I concetti fondamentali del metodo rappresentativo* in « I problemi del servizio sociale », maggio-giugno 1952, nei quali vengono richiamati i principali lavori di C. Gini, F. Yates, L.H. Tippet, G. Pompilj, V. Castellano.

gono ogni volta 5 numeri a sorte dal complesso dei primi 90 numeri della serie naturale, senza rimettere l'elemento estratto nella urna. Conoscendo la massa dalla quale si estraggono i campioni possiamo calcolare esattamente, mediante la (18) o la (18''), la varianza delle differenze medie riguardanti i $\binom{90}{5} = 43.949.268$ campioni.

Essendo i numeri costituenti la massa in progressione aritmetica di ragione 1, si ha, com'è noto ([16]).

$$\Delta = \frac{H+1}{3}, \quad \Delta_R = \frac{H^2-1}{3H}$$
$$\sigma = \sqrt{\frac{H^2-1}{12}} \tag{34}$$

mentre per il calcolo di Δ_{21}, dalla sua espressione (19) si ricava, dopo facili trasformazioni, nel caso di numeri interi consecutivi,

$$\Delta_{21} = \frac{6}{H(H-1)(H-2)} \sum_{i=1}^{H} \frac{(i-1)\,i}{2} \cdot \frac{(H-i)(H-i+1)}{2} =$$
$$= \frac{(H+1)(H+2)}{20} \tag{35}$$

e

$$\Delta_{21,R} = \frac{(H^2-1)(H^2-4)}{20H^2}. \tag{35'}$$

Sostituendo questi valori nella (18) — o nella (18'') — e nella (20) si ha, rispettivamente :

$$Var(\mathbf{D}) = \frac{(H-N)(H+1)(N+3)}{45N(N-1)} \tag{36}$$

$$Var(\mathbf{D})_r = \frac{(H^2-1)[H^2(N+3)-4N+18]}{45H^2N(N-1)}. \tag{37}$$

Pertanto, nel caso di una v.c. discreta associata ad una massa finita di H numeri interi consecutivi, ciascuno dei quali abbia

([16]) C. Gini, *Variabilità e concentrazione*, op. cit., 2ª edizione, 1955, pag. 265.

probabilità $\frac{1}{H}$ di essere estratto, la varianza dei campioni di N termini è espressa dalla (36) o dalla (37) secondo che i campioni siano stati ottenuti seguendo lo schema dell'estrazione in blocco o quello dell'estrazione con ripetizione.

Nel caso particolare del gioco del lotto, $H = 90$, $N = 5$ e quindi si hanno i seguenti valori numerici:

$$\begin{aligned} &\Delta = 30{,}333, \quad \Delta_R = 29{,}996, \quad \sigma = 25{,}980, \\ &\Delta_{21} = 418{,}600, \quad \Delta_{21,R} = 404{,}750, \\ &Var\,(\mathbf{D}) = 68{,}7555, \quad \sigma_{\mathbf{D}} = 8{,}292 \\ &Var\,(\mathbf{D})_r = 71{,}9889, \quad \sigma_{\mathbf{D}_r} = 8{,}484 \end{aligned} \tag{38}$$

Il minor valore ottenuto applicando la (36) invece della (37) conferma che lo schema dell'estrazione in blocco fornisce, a parità di prove, una maggiore precisione dei risultati, nel senso che questi sono più addensati intorno al valore medio esatto.

Per accertare il grado di approssimazione a cui conducono la (23) e la (22'), si osservi che da esse si ottiene, rispettivamente,

$$\begin{aligned} &Var\,(\mathbf{D}) \simeq \frac{85}{90}\,71{,}9889 = 67{,}9895 \text{ e } \sigma_{\mathbf{D}} = 8{,}246 \\ &Var\,(\mathbf{D}) \simeq 1{,}01044\,(71{,}9889 - 8{,}9989) = 63{,}648 \text{ e } \sigma_{\mathbf{D}} = 7{,}978 \end{aligned} \tag{39}$$

La (23) porta dunque a risultati assai prossimi a quelli esatti. A queste formule si può ricorrere tutte le volte che si sia già calcolato la $Var\,(\mathbf{D})_r$ e si ritenga sufficiente calcolare approssimativamente il valore di $Var\,(\mathbf{D})$. [17]

---

[17] La variabile casuale discreta associata al gioco del lotto può essere approssimata, evidentemente, dalla variabile casuale continua con legge di probabilità espressa dalla distribuzione rettangolare di densità:

$$d\,F = \frac{1}{H}\,d\,x \qquad \text{con } H = 90\,.$$

Per tale distribuzione si ha (NAIR, 1936, op. cit.).

$$M\,(\mathbf{D}) = \frac{1}{3}\,H = 30$$

$$Var\,(\mathbf{D}) = \frac{H^2}{9}\,\frac{N+3}{5\,N\,(N-1)} = 72 \tag{a}$$

I precedenti risultati presuppongono la conoscenza dei valori riguardanti la massa. Essi risolvono il problema di probabilità diretta riguardante l'errore quadratico medio della differenza media. Ma alla fine del n. 5 abbiamo già accennato che, nelle applicazioni ordinarie della varianza delle costanti statistiche, alle quantità riguardanti la massa dei casi da cui i campioni sono tratti si sostituiscono le corrispondenti quantità stimate. Per vedere in concreto l'influenza di questa sostituzione, consideriamo le 530 cinquine estratte nel gioco del lotto nel 1954. Per ognuna di queste cinquine abbiamo calcolato la differenza media semplice senza ripetizione ${}_c\Delta$ e, inoltre, ${}_c\Delta_{21}$ con la (19) posto $H = 5$, e ${}_c\sigma^2$ con la (33).

La media dei 530 valori di ${}_c\Delta$ è risultata $= 30{,}201$ leggermente inferiore al valore teorico 30,333, mentre per la varianza e lo scarto quadratico medio si è ottenuto

$$Var\,({}_c\Delta) = \frac{1}{529}\left[\Sigma\,{}_c\Delta^2 - \frac{1}{530}\,(\Sigma\,{}_c\Delta)^2\right] = 70{,}775 \quad \text{e} \quad \sigma_{{}_c\mathbf{D}} = 8{,}413 \qquad (40)$$

---

per $N = 5$. Questi due risultati teorici sono da considerare, come si è detto, approssimati a quelli da noi calcolati, per il motivo detto di riferirsi alla v.c. continua invece che a quella discreta ; inoltre, essi si riferiscono ovviamente all'estrazione con ripetizione — e quindi a Δ_R — e non all'estrazione senza ripetizione, come sono generalmente ricavati i campioni nella pratica statistica.

Le (36) e (37) potrebbero essere scritte, rispettivamente

$$Var\,(\mathbf{D}) = \frac{H^2\,(N+3)}{45\,N\,(N-1)} - \frac{H\,(N+3)}{45\,N} - \frac{N+3}{45\,(N-1)} \qquad (36')$$

$$Var\,(\mathbf{D})_r = \frac{H^2\,(N+3)}{45\,N\,(N-1)} - \frac{1}{9\,(N-1)} + \frac{1}{3\,N\,(N-1)} + \frac{4\,N-18}{45\,H^2\,N\,(N-1)} \qquad (37')$$

nelle quali il primo termine corrisponde al risultato del Nair, e i termini aggiuntivi esprimono le modifiche da apportare a questo per ottenere la varianza della differenza media dei campioni di N elementi ottenuti da una massa finita di H numeri interi consecutivi, mediante estrazione in blocco o, rispettivamente, con estrazione bernoulliana.

I termini aggiuntivi della (37') sono generalmente di scarsa entità e possono essere trascurati in molti casi pratici. Non altrettanto si può dire dei termini aggiuntivi della (36'), la cui entità dipende dalle dimensioni della massa e dei campioni.

che differiscono per eccesso da quelli esatti indicati nel penultimo rigo delle (38). Le differenze possono considerarsi del tutto casuali e la loro entità può benissimo giustificarsi tenendo conto che i campioni considerati sono appena 530 su 43.949.268 pari allo 0.012‰ del totale di tutte le possibili cinquine.

Questa stessa circostanza rende poco significativa la distribuzione dei valori calcolati per i $_c\Delta$. Tuttavia mi pare interessante, come primo saggio, indicare la distribuzione che risulta dai valori in esame, salvo a ritornare su questo argomento in una prossima occasione, avendo già determinato la distribuzione teorica dei $_c\Delta$ per variabili casuali discrete che assumono H valori equidistanti, tutti con la stessa probabilità.

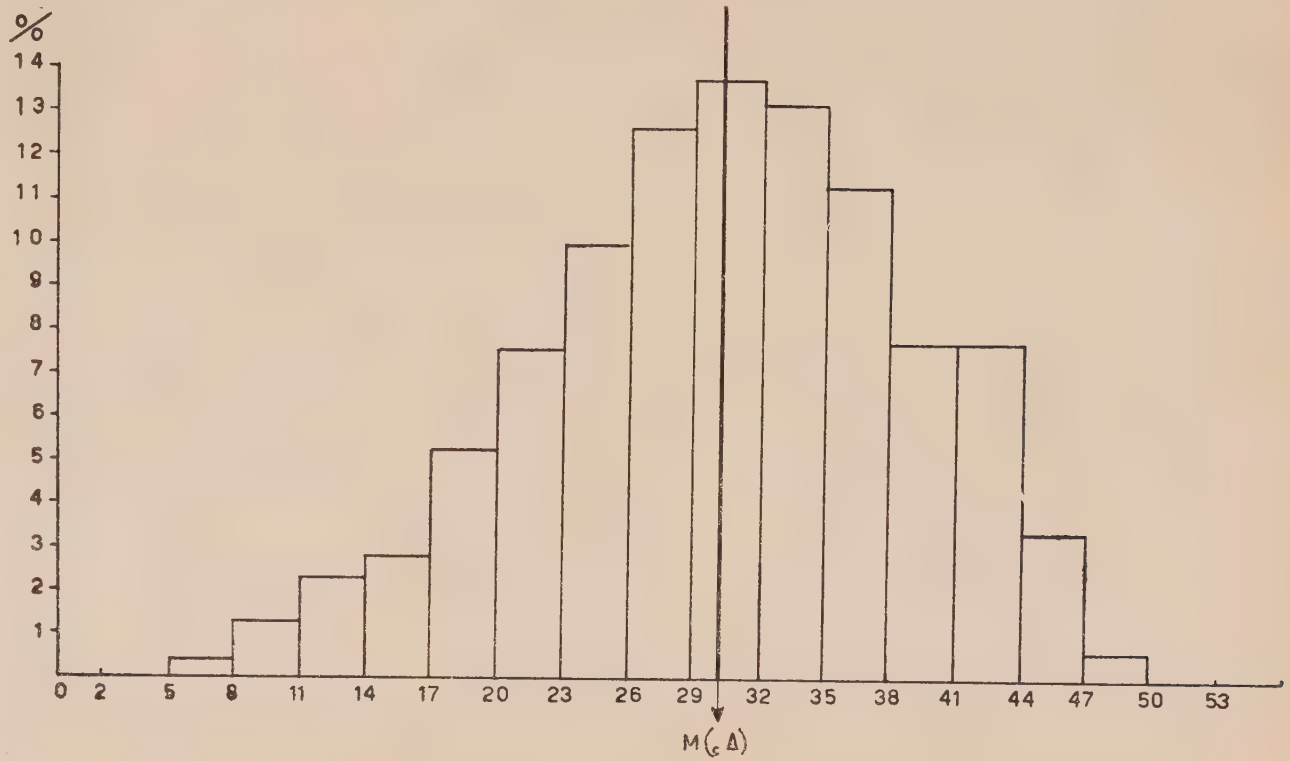
Tanto nella tav. 1 che nel grafico — il cui istogramma presenta una chiara asimmetria — i $_c\Delta$ sono stati raggruppati in classi di 3 unità ciascuno, a partire dal minimo valore possibile che è uguale a 2 (caso di 5 numeri interi consecutivi) fino al massimo che è 53 (nel caso che i primi 2 numeri siano 1 e 2 e gli ultimi due siano 89 e 90).

TAVOLA 1.

Distribuzione della differenza media delle cinquine estratte al gioco del lotto nell'anno 1954

Classi di valori per $_c\Delta$	Frequenze	
	assolute	percentuali
5 |– 8	2	0,38
8 |– 11	7	1,32
11 |– 14	12	2,26
14 |– 17	15	2,83
17 |– 20	28	5,28
20 |– 23	40	7,55
23 |– 26	53	10,00
26 |– 29	67	12,64
29 |– 32	73	13,77
32 |– 35	70	13,21
35 |– 38	60	11,32
38 |– 41	41	7,74
41 |– 44	41	7,74
44 |– 47	18	3,40
47 |– 50	3	0,56
IN COMPLESSO	530	100,00

GRAFICO 1. — *Distribuzione della differenza media delle cinquine estratte al gioco del lotto nell'anno* 1954.



Nella totalità dei 530 campioni considerati, tra $M({}_c\Delta) \pm \pm \sqrt{Var({}_c\Delta)} = 30{,}201 \pm 8{,}413$, cadono approssimativamente 354 valori ${}_c\Delta$, pari al 66,8 %, mentre tra $30{,}201 \pm 2.8{;}413$ cadono il 95,9 % dei casi.

Questi risultati, come si è detto, hanno un valore di prima approssimazione del problema riguardante la distribuzione teorica dei ${}_c\Delta$, in attesa di più approfondita indagine al riguardo.

Vediamo ora cosa succede se nella (18) sostituiamo ai vari parametri le corrispondenti quantità stimate su un singolo campione. I risultati che si ottengono sono molto discordanti tra loro e spesso alquanto lontani dai valori esatti; financo dei risultati negativi in alcuni casi! [18]

Ciò dipende ovviamente dal ben noto fatto che la sostituzione detta non è lecita, e si usa ammetterla quando si tratta di campioni sufficientemente rappresentativi della massa per quanto riguarda i parametri da utilizzare; e poichè tale ipotesi non è fa-

---

[18] Ad esempio, per il seguente campione 76, 20, 9, 73, 28, (estrazione di Bari del 26 giugno 1954) si ha: ${}_c\Delta = 37{,}4$, ${}_c\Delta_{21} = 428{,}4$, ${}_c\sigma^2 = 959{,}91$ Sostituendo questi valori nella (18) si ha $Var(\mathbf{D}) = -29{,}28$.

Tale risultato inaccettabile mi ha indotto a fare alcune considerazioni supplementari. Ripetendo, ad esempio, ogni numero del precedente campione 18 volte (in modo da formare una massa di 90 numeri) la varianza della differenza media di questa nuova massa calcolata con la (18) è ancora negativa?

Ora i parametri (che distingueremo dai precedenti ponendo un'apice) di questa nuova massa si ricavano dai precedenti in base alle seguenti relazioni:

$$\Delta' = \frac{90}{89} \cdot \frac{4}{5}\, {}_c\Delta = 30{,}256$$

$$\Delta'_{21} = \frac{90^2}{89 \cdot 88} \cdot \frac{4 \cdot 3}{5^2}\, {}_c\Delta_{21} = 212{,}668$$

$$\sigma'^2 = \frac{18 \cdot 4}{89}\, {}_c\sigma^2 = 776{,}556.$$

Sostituendo questi nuovi valori nella (18) si ricava

$$Var(\mathbf{D})' = 72{,}28$$

giungendo così a un valore positivo. I parametri di un solo campione, potendo dunque divergere sensibilmente dai valori della massa, con entità affatto proporzionali per tutti essi, possono determinare l'inconveniente segnalato quando vengono sostituiti nella (18).

cilmente controllabile in molti casi concreti, spesso la si ammette tacitamente o soltanto con qualche esplicita riserva.

Nel nostro caso i campioni sono cinquine e non possiamo aumentare il numero degli elementi costitutivi di ciascuno di essi per renderli più rappresentativi. Ne consegue che con campioni così piccoli è molto azzardato ricorrere alla sostituzione in parola per fare inferenze sull'intervallo di fiducia entro il quale, con una data probabilità, cade il valore della differenza media della massa. D'altra parte, è nota l'importanza dei piccoli campioni, specialmente nel campo sperimentale.

Essendo N fisso può essere quindi utile indagare sui risultati che si ottengono sostituendo ai singoli parametri la media delle 53 estrazioni avvenute nelle singole città dove si fanno le estrazioni del lotto, oppure sostituendo la media di tutte le 530 estrazioni nel lotto del 1954. I risultati sono riportati nella tavola 2.

TAVOLA 2.

Valori medi di ${}_c\Delta$, ${}_c\Delta_{21}$, ${}_c\sigma^2$ *e valori delle varianze effettive nonchè di quelle dedotte dalla* (18) *riguardanti le cinquine estratte nelle singole ruote e nel complesso delle ruote nell'anno* 1954.

Città di estrazione (Ruota)	$M({}_c\Delta)$	$M({}_c\Delta_{21})$	$M({}_c\sigma^2)$	Varianza effettiva	Varianza dedotta con (18) $Var({}_c\Delta)$	$M({}_c\Delta) \pm 2\sqrt{Var({}_c\Delta)}$
Bari	29,72	410,22	647,28	58,84	68,74	13,14—46,30
Cagliari	30,95	433,23	682,04	51,65	54,41	16,18—45,70
Firenze	31,97	449,00	742,49	53,69	64,56	15,91—48,03
Genova	28,61	425,71	591,51	72,21	74,89	11,31—45,91
Milano	31,10	440,22	713,83	90,84	75,65	13,70—48,50
Napoli	31,81	464,71	730,05	73,59	67,42	15,39—48,23
Palermo	29,23	405,55	648,41	97,47	87,68	10,51—47,95
Roma	28,42	379,95	602,10	89,23	72,82	11,36—45,48
Torino	30,04	408,29	662,80	58,45	67,33	13,62—46,46
Venezia	30,16	423,09	660,67	59,78	66,46	13,86—46,46
COMPLESSO DELLE 530 ESTRAZIONI	30,20	424,00	668,12	70,78	70,96	13,36—47,04
PER LA MASSA	30,33	418,60	674,92	—	68,76	13,75—46,91

Le differenze tra i valori effettivi della varianza delle differenze medie osservate sulle cinquine di ciascuna « ruota » e quelli provenienti dalla (18) sono più elevate della differenza che si riscontra tra i corrispondenti valori riguardanti il complesso delle 530 estrazioni dell'anno. Nessuna sistematicità si riscontra, esaminando le differenze tra i valori dedotti con la (18) e il valore della varianza riguardante l'universo di tutti i campioni. Queste differenze, prese in valore assoluto, passano da un minimo di 0,02 ad un massimo di 18,96; il minimo non corrisponde alla varianza ottenuta con i valori medi dei 530 campioni. Nei casi singoli si può quindi verificare che aumentando il numero dei campioni diminuisca l'approssimazione, mentre è noto che, in media, aumentando il numero di questi, ci si avvicina di più al valore esatto.

Servendoci dei valori dedotti dalla (18), gli intervalli di fiducia riportati nell'ultima colonna della tav. 2 mostrano che in nessun caso si esce dai limiti effettivi dei ${}_c\Delta$ che, come si è detto, sono 2 e 53. A causa dell'asimmetria della distribuzione dei ${}_c\Delta$ i limiti superiori differiscono da 53 per meno di una volta $\sigma_{\mathbf{D}}$. Anche se avessimo aggiunto a $M({}_c\Delta)$ una sola volta l'errore quadratico medio avremmo avuto degli intervalli di fiducia contenenti il valore esatto della differenza media. Nel caso, dunque, che N sia fisso e di scarsa entità, fondandoci su parametri ricavati facendo la media delle stime riguardanti i singoli campioni si ottiene, come era da aspettarsi, un sensibile miglioramento per il calcolo dell'intervallo di fiducia della differenza media.

RIASSUNTO

La varianza della differenza media dei campioni è nota nel caso che questi siano ricavati secondo lo schema delle prove ripetute. L'Autore, invece, suppone che i campioni di N elementi siano ottenuti mediante estrazione a caso *senza ripetizione* da una massa finita di H termini. Tale schema probabilistico è il più aderente alla pratica statistica, ma, nello stesso tempo, è più complesso del precedente perchè le singole prove non sono indipendenti.

Il procedimento seguito dall'A. è fondato su una opportuna trasformazione della v.c. (variabile casuale) **D** descritta dalle stime ${}_c\Delta$ della differenza

media riguardanti i singoli campioni deducibili dalla massa data. Il risultato ottenuto esprime la varianza di **D** in funzione di H, di N e delle seguenti costanti caratteristiche della massa: differenza media, scarto quadratico medio e differenza media di ordine 2 espressa dalla (19).

Questo risultato, confrontato con quello valido nel caso dello schema bernoulliano, mostra che quest'ultimo si ottiene dal nostro mediante passaggio al limite per H tendente all'infinito. La (18), quindi, generalizza il risultato degli Autori che si sono occupati in precedenza dell'argomento.

Poichè nella pratica statistica solo raramente si dispone delle costanti riguardanti la massa, l'A., per esaminare l'influenza che sui valori della varianza di **D** ha la sostituzione di dette costanti con le corrispondenti quantità stimate su singoli campioni, prende in considerazione le cinquine tratte dal gioco del lotto in Italia. Per tale gioco è evidentemente possibile calcolare tanto i valori delle costanti statistiche riguardanti la massa, quanto i valori riguardanti i singoli campioni. Mostra che, quando N è piccolo, la detta sostituzione può portare a valori inaccettabili; in tal caso e se i campioni hanno dimensioni prestabilite, la sostituzione di dette costanti, con valori medi calcolati su un numero crescente di campioni, migliora notevolmente la stima della varianza della differenza media.

RÉSUMÉ

La variance de la différence moyenne des échantillons est connue dans le cas que ceux-ci soient tirés suivant le schéma des épreuves répétées. L'auteur du présent article suppose, au contraire, que les échantillons de N éléments soient obtenus, par tirage au hasard *sans répétition*, d'une masse détérminée de H éléments. Ce schéma des probabilités est celui qui est le plus conforme à la pratique statistique, mais il est, en même temps, plus compliqué que le précédent, parce que chaque épreuve n'est pas indépendante des autres.

Le procédé suivi par l'auteur est fondé sur une transformation appropriée de la v.a. (variable aléatoire) **D** décrite par les estimations $_c\Delta$ de la différence moyenne concernants les échantillons individuels que l'on peut tirer de la masse donnée. Le résultat obtenu exprime la variance de **D** en fonction de H, de N et des suivantes caractéristiques constantes de la masse: différence moyenne, écart quadratique moyen et différence moyenne de l'ordre 2 exprimée par la (19).

En outre, ce résultat, comparé à celui qui est valable dans le cas du schéma de Bernoulli, fait voir que ce dernier résultat s'obtient du nôtre au moyen du passage à la limite par H tendant à l'infini. La (18) généralise donc le résultat des auteurs qui se sont précédemment occupés de cette question.

Puisque, dans la réalité statistique, on ne dispose que rarement des constantes concernant la masse, l'auteur, pour examiner l'influence qu'exerce,

sur les valeurs de la variance de **D**, la substitution de ces constantes par les quantités correspondantes estimées sur des échantillons individuels, prend en considération les quines sorties au jeu de la loterie en Italie. Pour ce jeu, il est évidemment possible de calculer aussi bien les valeurs des constantes statistiques concernant la masse que les valeurs concernant des échantillons individuels. L'auteur montre que, lorsque N est bas, cette substitution peut porter à des valeurs inacceptables ; dans ce cas et si les échantillons consistent d'un nombre d'éléments préétabli, la substitution des dites constantes par des valeurs moyennes, calculées sur un nombre croissant d'échantillons, améliore notablement l'estimation de la variance de la différence moyenne.

SUMMARY

The variance of the mean difference of samples is known in the case that samples are drawn according to the process of repeated trials. The writer however supposes that samples of N elements are obtained by means of random selection *without replacement* from a finite universe of H terms This probability scheme is very closely connected to statistical practices, but at the same time it is more complex than the former, because the single trials are not independent.

The procedure followed by the writer is based on a suitable transformation of the random variable **D** expressed by the estimates of the mean difference of the single samples deducible from the given universe.

The result obtained expresses the variance of **D** in function of H and N and of the following parameters of the universe : mean difference, standard deviation, mean difference of second order expressed by (19).

Furthermore, such a result compared with the valid results as in the case of the Bernoulli scheme, shows that the latter is obtained from the former, by means of passing to the limit as H approaches infinity. The expression (18) therefore generalises the result of authours who have previously dealt with the subject.

Since in statistical reality parameters of Parents population are only rarely available, the writer, in order to examine what influence the substitution of the said parameters by the corresponding estimated quantities from single samples, has on the values of the variance of **D**, considers the five numbers extracted in the Italian lottery.

Obviously in this case it is possible to calculate both the values of the statistical parameters of the universe and the values of single samples. This shows that when N is small the said substitution can lead to values which are unacceptable ; in this case, and if the samples are of a fixed size the substitution of the said parameters by mean values calculated from an increasing number of samples, improves considerably the estimate of the variance in the mean difference.

ZUSAMMENFASSUNG.

Die Streuung der mittleren Differenz der Stichproben ist in dem Falle bekannt, in welchem diese letzten nach dem Verfahren der Zurücklegung erhalten werden. Der Verfasser, dagegen, nimmt an, dass die Stichproben von N Elementen durch Auslosung « ohne Zurücklegen » von einer begrenzten Masse von H Elementen erhalten werden. Dieses Wahrscheinlichkeitsschema ist in der statistischen Praxis zutreffender doch komplizierter als das vorhergehende, da die einzelnen Proben nicht unabhängig sind.

Das vom Verfasser verfolgte Verfahren beruht auf einer günstigen Umbildung der v.c. (der zufälligen Variable) **D** bestimmt durch die Scatzungswerte $_c\Delta$ der mittleren Differenz, welche die von der gegebenen Gesammtheit ableitbare Stichproben betreffen. Das erlangte Ergebnis drückt die Streuung von D als Funktion von H aus und auch von N und den folgenden Parametern : mittlere Differenz, mittlere quadratische Abweichung und mittlere Differenz zweiten Grades, von der Formel (19) ausgedrückt.

Wenn man dieses Ergebnis mit dem, welches im Bernoulli's Schema gültig ist, vergleicht, so ergibt sich, dass man das letzte von dem ersten erhält indem man für H unendlich gross, zur Grenze übergeht. Die Formel (18) verallgemeinert also das Ergebnis jener Verfasser die vorher dieses Argument behandelt haben.

Da man in der statistischen Praxis nur selten über die Parameter der Gesammtheit verfügt, nimmt der Verfasser das Lottospiel in Betracht um zu erfahren welchen Einfluss die Substitution dieser Parameter durch die entsprechenden Scätzungswerten der einzelnen Stichproben auf die Werte der Streuung **D** hat. In diesem Spiele ist es möglich sowohl die Parameter der Grundgesammtheit als auch die Werte die die einzelnen Stichproben betreffen zu berechnen. Das Resultat ergibt, dass, wenn N sehr klein ist, die obgenannte Substitution zu unannehmbaren Werten führen kann ; in diesem Falle bessert die Substitution solcher Konstanten durch mittlere Werte, berechnet mittelst einer zunehmenden Zahl von Stichproben, in bedeutender Weise die Scätzung der Streuung der mittleren Differenz.

CARLO BENEDETTI

Di un massimo dell'indice quadratico di oscillazione (1)

L'indice quadratico di oscillazione di una successione di reali positivi $b_1, b_2, \ldots b_n$ è:

$$^2O = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (b_i - b_{i+1})^2} \qquad (2)$$

che varia, in generale, in corrispondenza delle varie successioni-permutazioni di $b_1, b_2, \ldots b_n$. In questo lavoro verrà determinato il valore massimo che ^2O raggiunge al variare negli $n!$ modi possibili l'ordine di $b_1, b_2, \ldots b_n$.

Dalla determinazione del massimo, risulterà pure determinato il minimo fra gli $n!$ valori suddetti nonchè le distribuzioni massimanti e minimanti.

Verrà dimostrato il seguente teorema:

TEOREMA — *Data la successione non decrescente di reali positivi non tutti uguali*:

$$a_1 \leq a_2 \leq \ldots \leq a_n \qquad (1)$$

(1) Questo lavoro è stato oggetto di una comunicazione al Seminario di Statistica della Facoltà di Scienze Statistiche Dem. e Attuariali della Università di Roma il 16 Aprile 1955.

(2) Gli anglosassoni chiamano questo indice *mean square successive difference*. La radice s'intende presa sempre non negativa.

il massimo valore che ^2O *può assumere in corrispondenza delle* $n!$ *successioni-permutazioni della* (1) *è dato da:*

$$max\,^2O = \sqrt{\frac{\sum_{i=1}^{n-1} \lambda_i (a_i - a_{i+1})^2 + 2 \sum_{i<j} \mu_{ij} (a_i - a_{i+1}) (a_j - a_{j+1})}{n-1}} \qquad (2)$$

dove:

per n *pari*

$$\begin{cases} \lambda_i = n - |\,2i - n\,| \;\; per \;\; i \neq \frac{n}{2}\,; \;\; \lambda_{\frac{n}{2}} = n - 1 \\ \\ \mu_{ij} = i + n - j - |\,i - n + j\,| \;\; per \;\; i \neq n - j\,; \\ \mu_{ij} = 2i - 1 = 2(n-j) - 1 \;\; per \;\; i = n - j \end{cases}$$

per n *dispari*

$$\begin{cases} \lambda_i = n - |\,2i - n\,| \;\; per \;\; i \neq \frac{n-1}{2}, \frac{n+1}{2}\,; \\ \\ \begin{cases} \lambda_{\frac{n-1}{2}} = n - 1\,; \lambda_{\frac{n+1}{2}} = n - 2 \;\; se \;\; a_{\frac{n+1}{2}} \geq \frac{1}{2}\left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right) \\ oppure \\ \lambda_{\frac{n-1}{2}} = n - 2\,; \lambda_{\frac{n+1}{2}} = n - 1 \;\; se \;\; a_{\frac{n+1}{2}} \leq \frac{1}{2}\left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right) \end{cases} \\ \\ \mu_{ij} = i + n - j - |\,i - n + j\,| \;\; per \;\; i \neq n - j\,; \\ \mu_{ij} = 2i - 1 = 2(n-j) - 1 \;\; per \;\; i = n - j\,; \end{cases} \qquad \sum_{i<j} = \sum_{j=2}^{n-1} \sum_{i=1}^{j-1}$$

La dimostrazione di questo teorema verrà articolata attraverso quelle di alcuni lemmi.

Lemma 1 — *Se* $b_1, b_2, \ldots b_n$ *è una qualunque delle* $n!$ *successioni-permutazioni della* (1) *sarà:*

$$(n-1)\,^2O^2(b) = \sum_{i=1}^{n-1} (b_i - b_{i+1})^2 = \sum_{i=1}^{n-1} \lambda_i (a_i - a_{i+1})^2 + 2 \sum_{i<j} \mu_{ij} (a_i - a_{i+1}) (a_j - a_{j+1}) \qquad (3)$$

con $\lambda_i \geq 1\,;\; \mu_{ij} \geq 0\,.$

DIMOSTRAZIONE — Evidentemente sarà :

$$| b_h - b_{h+1} | = a_{r+h} - a_r = a_{r+h} - a_{r+h-1} + a_{r+h-1} - a_{r+h-2} + \ldots$$
$$\ldots + a_{r+1} - a_r \qquad (k \geq 1) \tag{4}$$

quindi λ_i è il numero degli intervalli del tipo $| b_h , b_{h+1} |$ (indicheremo così l'intervallo di lunghezza $| b_h - b_{h+1} |$ considerato sempre nel senso positivo) che contengono l'intervallo (a_i , a_{i+1}), intendendo dire che un intervallo $| b_h , b_{h+1} | = (a_r , a_{r+k})$ con $k \geq 1$ contiene l'intervallo (a_s , a_t) $(t > s)$ se $r \leq s ; r + k \geq t$.

Analogamente μ_{ij} è il numero degli intervalli $| b_h , b_{h+1} |$ che contengono ciascuno ambedue gli intervalli (a_i , a_{i+1}) e (a_j , a_{j+1}) $[i < j]$; quindi applicando la (4) ad ogni intervallo $| b_h , b_{h+1} |$ si ha immediatamente la (3) dove $\lambda_i \geq 1$ e $\mu_{ij} \geq 0$ poichè se vi fosse un $\lambda_i < 1$ ciò significherebbe che $\lambda_i = 0$ (non potrà essere mai ovviamente $\lambda_i < 0$) cioè che l'intervallo (a_i , a_{i+1}) non è mai contenuto in nessuno degli intervalli $| b_h , b_{h+1} |$ il che equivale ad ammettere che nella successione $b_1 , b_2 , \ldots b_n$ non esistono due termini b_h, b_{h+1} tali che uno dei due sia un a_r con $r \leq i$ e l'altro sia un a_s con $s \geq i + 1$; ma in tal caso vi sarebbero due successioni la prima di i termini a_r con $r \leq i$ e l'altra di $n - i$ termini a_s con $s \geq i + 1$ da non potersi unire in una successione di n termini in quanto ogni termine a_r $(r \leq i)$ non potrebbe essere che adiacente a termini $a_{r'}$ con $r' \leq i$ ed ogni termine $a_s (s \geq i+1)$ non essere che adiacente a termini $a_{s'}$, $(s' \geq i + 1)$ contro l'ipotesi della successione unica $b_1 , b_2 , \ldots b_n$ di n termini da cui siamo partiti.

In quanto a μ_{ij} è chiaro che due intervalli (a_i , a_{i+1}) e (a_j , a_{j+1}) o sono ambedue contenuti in un intervallo $| b_h , b_{h+1} |$ o non lo sono (non lo potranno essere comunque negativamente) ; quindi $\mu_{ij} \geq 0$.

Ponendo in (3) $\lambda_i = 1$; $\mu_{ij} = 0$ per ogni i , j si avrà per ${}^2O\,(b)$ un valore ω tale del quale non potrà essere inferiore nessuno degli $n !$ valori di ${}^2O\,(b)$ ottenuti in corrispondenza delle $n !$ successioni-permutazioni della (1). D'altra parte si vede subito che ω è l'indice quadratico di oscillazione ${}^2O\,(a)$ della (1) che è così la successione minimante ; quindi $\omega = \min {}^2O$ fra tutti gli $n !$ valori di ${}^2O\,(b)$. Si ha cioè il

COROLLARIO 1 — *Se* $a_1 \leq a_2 \leq \ldots \leq a_n$ *è una successione non decrescente di reali positivi, si ha, al permutare negli* $n!$ *modi possibili le* a_i

$$min\ ^2O = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (a_i - a_{i+1})^2}$$

n PARI

LEMMA 2 — *Se* $i \neq \frac{n}{2}$ *si ha:*

$$\lambda_i \leq n - |2i - n|$$

DIMOSTRAZIONE — Gli intervalli $|b_h, b_{h+1}|$ di cui $(n-1)\,^2O^2(b)$ costituisce la somma dei quadrati delle loro misure sono in numero di $n-1$ ed all'infuori di b_1 e b_n ogni b_i compare come estremo in due degli $n-1$ intervalli; b_1 e b_n compaiono invece una sola volta rispettivamente in $|b_1, b_2|$ e $|b_{n-1}, b_n|$.

Ora gli intervalli $|b_h, b_{h+1}| = (a_r, a_{r+k})$ che contengono (a_i, a_{i+1}) dovranno avere $r \leq i$; $r + k \geq i + 1$ e se $i < n - i$ per quello che è stato detto il numero degli intervalli $|b_h, b_{h+1}|$ contenenti (a_i, a_{i+1}) non può superare $2i$; quindi in tal caso si ha $\lambda_i \leq 2i$. Analogamente se $i > n - i$ sarà $\lambda_i \leq 2(n-i)$ quindi:

$$\lambda_i \leq min\ [2i, 2(n-i)] = n - |2i - n|\ (^1)$$

LEMMA 3: $\lambda_{\frac{n}{2}} \leq n - 1$

---

(1) Come si vede subito, in un intervallo reale (a, b), il minimo tra due funzioni reali e limitate $f(x)$ e $g(x)$ in (a, b) è dato da:

$$min\ [f(x), g(x)] = \frac{1}{2}\,[f(x) + g(x) - |f(x) - g(x)|]$$

DIMOSTRAZIONE — Se b_1 e b_n sono costituiti da a_r, a_{r+k} con $r, r+k \leq \frac{n}{2}$ il numero degli intervalli $|b_h, b_{h+1}|$ contenenti l'intervallo $(a_{\frac{n}{2}}, a_{\frac{n}{2}+1})$ non può superare $n-2$ poichè il numero degli estremi sinistri degli intervalli $|b_h, b_{h+1}|$ contenenti $(a_{\frac{n}{2}}, a_{\frac{n}{2}+1})$ non può, per ciò che è già stato detto nella dimostrazione del lemma precedente, superare $2\frac{n}{2} - 2 = n-2$. Se $r, r+k \geq \frac{n}{2} + 1$ si arriva lo stesso ad $n-2$ considerando il numero degli estremi destri degli intervalli $|b_h, b_{h+1}|$ contenenti $\left(a_{\frac{n}{2}}, a_{\frac{n}{2}+1}\right)$. Se $r \leq \frac{n}{2}$, $r+k \geq \frac{n}{2} + 1$ si vede analogamente che il numero degli intervalli $|b_h, b_{h+1}|$ contenenti $\left(a_{\frac{n}{2}}, a_{\frac{n}{2}+1}\right)$ non può superare $n-1$. Siccome si avrà certamente uno dei tre casi:

$$r, r+k \leq \frac{n}{2}\,; \quad r, r+k \geq \frac{n}{2} + 1\,; \quad r \leq \frac{n}{2}, \quad r+k \geq \frac{n}{2}$$

il lemma rimane dimostrato.

LEMMA 4 — *Se* $i \neq n-j$ *si ha*:

$$\mu_{ij} \leq i + n - j - |i - n + j|$$

DIMOSTRAZIONE: μ_{ij} è il numero degli intervalli $|b_h, b_{h+1}|$ che contengono ciascuno la coppia di intervalli (a_i, a_{i+1}); (a_j, a_{j+1}). Evidentemente l'estremo sinistro di ognuno di tali intervalli $|b_h, b_{h+1}|$ sarà un a_r con r non superiore ad i mentre l'estremo destro sarà un a_s con s non inferiore ad $j+1$; ora se $i < n-j$ il numero di tali intervalli $|b_h, b_{h+1}|$ non potrà superare $2i$ non potendo ogni estremo b_h o b_{h+1} comparire in più di due intervalli. Analogamente se $i > n-j$ il numero di tali intervalli $|b_h, b_{h+1}|$ non potrà superare $2(n-j)$ quindi:

$$\mu_{ij} \leq min\,[2i, 2(n-j)] = i + n - j - |i - n + j|$$

LEMMA 5 — *Se* $i = n - j$ *si ha*:

$$\mu_{ij} \leqq 2\,i - 1 = 2\,(n - j) - 1$$

DIMOSTRAZIONE: μ_{ij} è il numero degli intervalli $|\,b_h\,,\,b_{h+1}\,|$ che contengono ciascuno la coppia di intervalli $(a_i\,,\,a_{i+1})$; $(a_j\,,\,a_{j+1})$. Ora è chiaro che $2\,i\left(i < \frac{n}{2}\right)$ intervalli del tipo $|\,b_h\,,\,b_{h+1}\,|$ presi senza ripetizione fra gli $n - 1$ intervalli $|b_1\,,\,b_2\,|$; $|\,b_2\,,\,b_3\,|\,,\ \ldots\ |\,b_{n-1}\,,\,b_n\,|$ contengono almeno $2\,i + 1$ estremi b_r con r sempre diverso poichè se questi $2\,i$ intervalli sono consecutivi contengono $2i + 1$ estremi b_r altrimenti ne contengono di più. Quindi $\mu_{i,n-i} < 2\,i$ poichè $2\,i$ intervalli includendo almeno $2\,i + 1$ estremi b_r di posto diverso includerebbero oltre ai $2\,i$ estremi (i estremi $b_r = a_s$ con $s \leq i$ ed i estremi $b_{r'} = a_{s'}$, con $s' \geq n - - i + 1$) non interni all'intervallo $(a_i\,,\,a_{j+1})$ anche un estremo interno al suddetto intervallo $(a_i\,,\,a_{j+1}) = (a_i\,,\,a_{n-i+1})$. Quindi $\mu_{i,n-i} \leqq 2\,i - 1 = 2\,(n - j) - 1$ in quanto $2\,i - 1$ intervalli del tipo $|\,b_h\,,\,b_{h+1}\,|$ possono avere solo come minimo il numero $2\,i$ di estremi b_r con r sempre diverso.

LEMMA 6 — *Se facciamo in* (3):

$$\lambda_i = n - |\,2\,i - n\,| \quad per \quad i \neq \frac{n}{2}\,; \quad \lambda_{\frac{n}{2}} = n - 1$$

$$\mu_{ij} = i + n - j - |\,i - n + j\,| \quad per \quad i \neq n - j\,;$$

$$\mu_{ij} = 2\,i - 1 = 2\,(n - j) - 1 \quad per \quad i = n - j$$

si ottiene un valore Ω *che non può essere superato da nessuno degli* $n!$ *valori di* $(n - 1)\ ^2O^2\ (b)$ *ottenuti permutando le* b_i.

DIMOSTRAZIONE — Basta osservare che ciascun λ_i e μ_{ij} scelto non può, per i lemmi precedenti, essere superato dai corrispondenti λ_i e μ_{ij} relativi agli $n!$ indici quadratici di oscillazione ottenuti permutando la (1).

LEMMA 7 — *L'indice quadratico di oscillazione delle due successioni-permutazioni della* (1) *formate ed ordinate come segue:*

$$\text{I}\quad \begin{cases} a_{\frac{n}{2}+r}\,,\ a_{\frac{n}{2}-r} & \left(r = 1\,,\ 3\,,\ 5\,,\ \ldots \frac{n}{2} - 1\right) \\ a_n & \\ a_{\frac{n}{2}-r}\,,\ a_{\frac{n}{2}+r} & \left(r = \frac{n}{2} - 2\,,\ \frac{n}{2} - 4\,,\ \ldots 4\,,\ 2\right) \\ a_{\frac{n}{2}} & \end{cases} \Bigg\} \textit{per } \frac{n}{2} \textit{ pari}$$

$$\text{II}\quad \begin{cases} a_{\frac{n}{2}+r}\,,\ a_{\frac{n}{2}-r} & \left(r = 1\,,\ 3\,,\ 5\,,\ \ldots \frac{n}{2} - 2\right) \\ a_n & \\ a_{\frac{n}{2}-r}\,,\ a_{\frac{n}{2}+r} & \left(r = \frac{n}{2} - 1\,,\ \frac{n}{2} - 3\,,\ \ldots 4\,,\ 2\right) \\ a_{\frac{n}{2}} & \end{cases} \Bigg\} \textit{per } \frac{n}{2} \textit{ dispari}$$

è uguale al valore $\sqrt{\dfrac{\Omega}{n-1}}$ *di cui al lemma precedente.*

DIMOSTRAZIONE — Anzitutto si nota che tanto nella (I) quanto nella (II) vi sono n termini appartenenti ad (1) e tutti di posto diverso quindi trattasi di due successioni-permutazioni della (1); inoltre si nota che tutte le coppie di termini consecutivi in (I) e (II) sono simmetriche negli indici rispetto ad $\frac{n}{2}$ oppure ad $\frac{n+1}{2}$ oppure ad $\frac{n}{2}+1$. (1)

---

(1) Cioè per ogni coppia $b_h\,,\ b_{h+1}$ si ha $\dfrac{h+h+1}{2} = \begin{cases} \frac{n}{2} \text{ oppure} \\ \frac{n+1}{2} \text{ oppure} \\ \frac{n}{2}+1 \end{cases}$

Inoltre è facile vedere che per la (I) e per la (II) si ha $\lambda_i = n - |2i - n|$ per $i \neq \frac{n}{2}$; infatti in tal caso ogni termine $a_r\ (r \leq i)$ se $i < n - i$ cioè se $i \leq \frac{n}{2} - 1$ [1] sta fra due termini $a_s, a_t \left(s, t \geq \frac{n}{2} + 1\right)$ e siccome sarà anche $i + 1 \leq \frac{n}{2}$ si avranno $2i$ intervalli $|b_h, b_{h+1}|$ contenenti (a_i, a_{i+1}); se invece $i > n - i$ cioè se $i + 1 \geq \frac{n}{2} + 2$ [1] ognuno degli $n - i$ termini $a_r\ (r \geq i + 1)$ sta fra due termini $a_s, a_t \left(s, t \leq \frac{n}{2}\right)$ e siccome sarà anche $i \geq \frac{n}{2} + 1$ si avranno $2\,(n-i)$ intervalli $|b_h, b_{h+1}|$ contenenti (a_i, a_{i+1}); quindi $\lambda_i = min\,[2i, 2\,(n-i)] = n - |2i - n|$ per $i \neq \frac{n}{2}$. Se $i = \frac{n}{2}$ si ha $\lambda_i = n - 1$ poichè ogni termine $a_r \left(r < \frac{n}{2}\right)$ sta tra due termini $a_s, a_t \left(s, t \geq \frac{n}{2} + 1\right)$ mentre $a_{\frac{n}{2}}$ fungendo da b_n [oppure da b_1 se invertiamo la intera successione-permutazione (I) o (II)] figura nel solo intervallo $|b_{n-1}, b_n|$ (oppure in $|b_1, b_2|$) e quindi vi saranno $2\left(\frac{n}{2} - 1\right) + 1 = n - 1$ intervalli $|b_h, b_{h+1}|$ contenenti $\left(a_{\frac{n}{2}}, a_{\frac{n}{2}+1}\right)$ cioè $\lambda_{\frac{n}{2}} = n - 1$.

Inoltre sia per la (I) che per la (II) è $\mu_{ij} = i + n - j - |i - n + j|$ se $i \neq n - j$ poichè se è $i < n - j$ si ha anche $i \leq \frac{n}{2} - 1$ [2] e quindi ogni termine $a_r\ (r \leq i)$ sta fra due termini $a_s, a_t\ (s, t \geq j + 1)$ in quanto se vi fosse qualche a_s od a_t

---

[1] Se $i < n - i$; $i \leq n - i - 1$; $i \leq \frac{n-1}{2}$; e, poichè n è pari, $i \leq \frac{n}{2} - 1$.

Se $i > n - i$; $i \geq n - i + 1$; $i \geq \frac{n+1}{2}$; $i \geq \frac{n}{2} + 1$; $i + 1 \geq \frac{n}{2} + 2$.

Se $i = n - i$; $i = \frac{n}{2}$.

[2] Poichè se $i < n - j$; $j < n - i$; $j \leq n - i - 1$; $i < n - i - 1$; $i \leq n - i - 2$; $i \leq \frac{n}{2} - 1$.

con $s\ ,\ t < j+1$ allora si troverebbero degli intervalli in (I) ed in (II) $|\ b_h\ ,\ b_{h+1}\ |$ con $\frac{h+h+1}{2} < \frac{n}{2}$ contro la proprietà di simmetria di cui godono le coppie di termini consecutivi nelle (I), (II) in quanto si avrebbe $\frac{h+h+1}{2} \leq \frac{i+n-i-1}{2} < \frac{n}{2}$ quindi sarà $\mu_{ij} = 2\,i$ se $i < n-j$. Con analogo ragionamento applicato agli $a_r\ (r \geq j+1)$ si dimostra che $\mu_{ij} = 2\,(n-j)$ se $i > n-j$ cioè se $j+1 \geq \frac{n}{2}+2$ [1] ed allora si avrà:

$$\mu_{ij} = min\,[2\,i\ ,\ 2\,(n-j)] = i+n-j-|\,i-n+j\,|$$

Infine per la (I) e (II) se $i = n-j$ cioè se $i < \frac{n}{2}$ [1] si ha $\mu_{ij} = 2i-1 = 2(n-j)-1$; infatti gli i termini $a_r (r \leq i)$ stanno ciascuno fra due termini $a_s\ ,\ a_t$ essendo questi ultimi in numero di $i+1$ e non potendo essere $s\ ,\ t < j$ per ragioni di simmetria (altrimenti si avrebbe un $|\ b_h\ ,\ b_{h+1}\ |$ con $\frac{h+h+1}{2} \leq \frac{i+n-i-1}{2} < \frac{n}{2}$) gli $i+1$ termini $a_s\ ,\ a_t$ dovranno essere $a_j\ ,\ a_{j+1}\ ,\ a_{j+2}\ ,\ \ldots\ a_n$ di cui non va considerato a_j in quanto interno [2] ad $(a_r\ ,\ a_{j+1})\ (r \leq i)$; ora, poichè solo a_i tra gli a_r con $r \leq i$ sarà, per le note proprietà di simmetria, adiacente ad a_j, avremo un intervallo di meno tra gli intervalli $|\ b_h\ ,\ b_{h+1}\ |$ che comprendono $(a_i\ ,\ a_{j+1})$ cioè $\mu_{ij} = \mu_{i,n-i} = 2i-1 = 2\,(n-j)-1$. Si hanno quindi per λ_i e per μ_{ij} tutti i valori che danno luogo a $\sqrt{\frac{\Omega}{n-1}}$ (vedi lemma 6).

Con la dimostrazione di questo ultimo lemma rimane dimostrata la prima parte del teorema (per n pari).

---

[1] Se $i > n-j\ ;\ i \geq n-j+1\ ;\ j > n-j+1\ ;\ j \geq n-j+2\ ;$

$$j \geq \frac{n}{2}+1\ ;\ j+1 \geq \frac{n}{2}+2.$$

Se $i = n-j\ ;\ i < n-i\ ;\ i < \frac{n}{2}$.

[2] Intendendo naturalmente sempre rispetto all'indice j come definito nella dimostrazione del lemma 1, dove viene data la definizione di un intervallo $|\ b_h\ ,\ b_{h+1}\ |$ che comprende un altro intervallo e quindi dei termini b_h.

LEMMA 8 — *Se* $i \neq \frac{n-1}{2}, \frac{n+1}{2}$ *si ha*:

$$\lambda_i \leq n - |2i - n|$$

DIMOSTRAZIONE — È analoga a quella del lemma 2.

LEMMA 9 — *Pur valendo per* $\lambda_{\frac{n-1}{2}}$ *e* $\lambda_{\frac{n+1}{2}}$ *separatamente la limitazione del lemma 8, se* $\lambda_{\frac{n-1}{2}} = n - 1$ *allora* $\lambda_{\frac{n+1}{2}}$ *non può superare* $n - 2$ *e viceversa se* $\lambda_{\frac{n+1}{2}} = n - 1$ *allora* $\lambda_{\frac{n-1}{2}}$ *non può superare* $n - 2$.

DIMOSTRAZIONE — Se $\lambda_{\frac{n-1}{2}} = n - 1$ si ha che l'intervallo $\left(a_{\frac{n-1}{2}}, a_{\frac{n+1}{2}}\right)$ è ripetuto $n - 1$ volte, cioè vi sono $n - 1$ intervalli $|b_1, b_{1+1}|$ contenenti $\left(a_{\frac{n-1}{2}}, a_{\frac{n+1}{2}}\right)$; non ve ne possono essere di più poichè (vedi dimostrazione lemma 2) ogni estremo $a_r \left(r \leq \frac{n-1}{2}\right)$ non può figurare in più di due intervalli $|b_h, b_{h+1}|$.

D'altra parte è ovvio che in tal caso b_1 e b_n sono costituiti da a_s e a_t con $s, t \geq \frac{n+1}{2}$; quindi vi saranno solo due intervalli $|b_1, b_{1+1}|$ contenenti rispettivamente a_s e a_t cioè b_1 b_n mentre per gli altri $a_r \left(r \geq \frac{n-1}{2}\right)$ vi sarà la solita limitazione di non comparire in più di due intervalli $|b_h, b_{h+1}|$. Concludendo se $\lambda_{\frac{n-1}{2}} = n - 1$ si avrà $\lambda_{\frac{n+1}{2}} \leq \left(n - \frac{n+1}{2}\right) 2 - 1 = n - 2$ in quanto b_1 e b_n sono rappresentati da a_s, a_t con $s, t \geq \frac{n-1}{2}$. Identica è la dimostrazione se $\lambda_{\frac{n+1}{2}} = n - 1$ per cui si ha $\lambda_{\frac{n-1}{2}} \leq$ $\leq \left(\frac{n-1}{2}\right) 2 - 1$ in quanto b_1 e b_n sono rappresentati da a_s, a_t con $s, t \leq \frac{n+1}{2}$.

LEMMA 10 — *Se* $i \neq n - j$ *si ha*

$$\mu_{ij} \leq i + n - j - |i - n + j|$$

DIMOSTRAZIONE — È analoga a quella del lemma 4.

LEMMA 11 — *Se* $i = n - j$ *si ha*:

$$\mu_{ij} \leq 2i - 1 = 2(n - j) - 1$$

DIMOSTRAZIONE — È analoga a quella del lemma 5.

LEMMA 12 — *Se facciamo in* (3):

$$\lambda_i = n - |2i - n| \text{ per } i \neq \frac{n-1}{2}, \frac{n+1}{2}$$

$$\mu_{ij} = n + i - j - |i - n + j| \text{ per } i \neq n - j \cdot$$

$$\mu_{ij} = 2i - 1 = 2(n - j) - 1 \text{ per } i = n - j$$

$$\cdot \begin{cases} \lambda_{\frac{n-1}{2}} = n - 1; \ \lambda_{\frac{n+1}{2}} = n - 2 \ \text{ se } \ a_{\frac{n+1}{2}} \geq \frac{1}{2}\left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right) \\ \lambda_{\frac{n-1}{2}} = n - 2; \ \lambda_{\frac{n+1}{2}} = n - 1 \ \text{ se } \ a_{\frac{n+1}{2}} \leq \frac{1}{2}\left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right) \end{cases}$$

si ottiene un valore Ω *che non può essere superato da nessuno degli* $n!$ *valori di* $(n - 1)\ {}^2O^2$ (b) *ottenuti permutando le* b_i.

DIMOSTRAZIONE — Basta osservare che ciascun λ_i e μ_{ij} scelto non può, per i lemmi precedenti, essere superato dai corrispondenti λ_i e μ_{ij} relativi agli $n!$ indici quadratici di oscillazione. È ovvio che per $\lambda_{\frac{n-1}{2}}$ e $\lambda_{\frac{n+1}{2}}$ verranno scelti quei due valori, fra le due coppie alternative, tali che si massimizzi $\lambda_{\frac{n-1}{2}} \Delta a_{\frac{n-1}{2}} + \lambda_{\frac{n+1}{2}} \Delta a_{\frac{n+1}{2}}$ cioè verrà scelto $\lambda_{\frac{n-1}{2}} = n - 1$ e $\lambda_{\frac{n+1}{2}} = n - 2$ se $\Delta a_{\frac{n-1}{2}} \geq \Delta a_{\frac{n+1}{2}}$ e $\lambda_{\frac{n-1}{2}} = n - 2$, $\lambda_{\frac{n+1}{2}} = n - 1$ se $\Delta a_{\frac{n-1}{2}} \leq \Delta a_{\frac{n+1}{2}}$.

LEMMA 13 — *L'indice quadratico di oscillazione delle successioni-permutazioni della* (1) *formate ed ordinate come segue:*

$$\left.\begin{array}{ll}
\text{III} & a_{\frac{n+1}{2}+r},\ a_{\frac{n-1}{2}-r} \quad \left(r = 1, 3, 5, \ldots \frac{n+1}{2} - 2\right) \\
\textit{per} & \\
\frac{n+1}{2} & a_n \\
\textit{dispari} & a_{\frac{n-1}{2}-r},\ a_{\frac{n+1}{2}+r} \quad \left(r = \frac{n+1}{2} - 3, \frac{n+1}{2} - 5, \ldots 2, 0\right) \\
\text{IV} & a_{\frac{n+1}{2}+r},\ a_{\frac{n-1}{2}-r} \quad \left(r = 1, 3, 5, \ldots \frac{n+1}{2} - 3\right) \\
\textit{per} & \\
\frac{n+1}{2} & a_n \\
\textit{pari} & a_{\frac{n-1}{2}-r},\ a_{\frac{n+1}{2}+r} \quad \left(r = \frac{n+1}{2} - 2, \frac{n+1}{2} - 4, \ldots 4, 2, 0\right)
\end{array}\right\} \textit{se } a_{\frac{n+1}{2}} \geq \frac{1}{2} \cdot \left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right)$$

$$\left.\begin{array}{ll}
\text{V} & a_{\frac{n+1}{2}-r},\ a_{\frac{n+3}{2}+r} \quad \left(r = 1, 3, 5, \ldots \frac{n+1}{2} - 2\right) \\
\textit{per} & \\
\frac{n+1}{2} & a_1 \\
\textit{dispari} & a_{\frac{n+3}{2}+r},\ a_{\frac{n+1}{2}-r} \quad \left(r = \frac{n+1}{2} - 3, \frac{n+1}{2} - 5, \ldots 2, 0\right) \\
\text{VI} & a_{\frac{n+1}{2}-r},\ a_{\frac{n+3}{2}+r} \quad \left(r = 1, 3, 5, \ldots \frac{n+1}{2} - 3\right) \\
\textit{per} & \\
\frac{n+1}{2} & a_1 \\
\textit{pari} & a_{\frac{n+3}{2}+r},\ a_{\frac{n+1}{2}-r} \quad \left(r = \frac{n+1}{2} - 2, \frac{n+1}{2} - 4, \ldots 2, 0\right)
\end{array}\right\} \textit{se } a_{\frac{n+1}{2}} \leq \frac{1}{2} \cdot \left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right)$$

è uguale al valore $\sqrt{\frac{\Omega}{n-1}}$ *di cui al lemma precedente.*

DIMOSTRAZIONE — Anzitutto si nota che in ciascuna delle 4 successioni vi sono n termini appartenenti ad (1) e tutti di posto diverso, quindi trattasi di 4 successioni-permutazioni della (1); inoltre si nota che in ciascuna delle suddette successioni tutte le coppie di termini consecutivi sono simmetriche negli indici

rispetto ad $\frac{n}{2}$ o ad $\frac{n+2}{2}$ o ad $\frac{n+1}{2}$.(1) Inoltre è facile vedere che in III e IV si ha $\lambda_i = n - |2i - n|$ per $i \neq \frac{n+1}{2}$; infatti in tal caso ogni termine $a_r\,(r \leq i)$ se $i < n - i$ cioè se $i \leq \frac{n-1}{2}$ (2) sta fra due termini $a_s, a_t \left(s, t \geq \frac{n+1}{2}\right)$ e siccome sarà anche $i + 1 \leq \frac{n+1}{2}$ si avranno $2i$ intervalli $|b_h, b_{h+1}|$ contenenti (a_i, a_{i+1}); se invece $i > n - i$ cioè se $i + 1 \geq \frac{n+5}{2}$ (2) ognuno degli $n - i$ termini $a_r\,(r \geq i + 1)$ sta fra due termini $a_s, a_t \left(s, t \leq \frac{n-1}{2}\right)$ e siccome sarà anche $i \geq \frac{n+3}{2}$ (2) si avranno $2(n - i)$ intervalli $|b_h, b_{h+1}|$ contenenti (a_i, a_{i+1}); quindi

$$\lambda_i = min\ [2i, 2(n - i)] = n - |2i - n|$$

Se $i = \frac{n+1}{2}$ si ha $\lambda_i = n - 2$ poichè ogni termine $a_r \left(r > \frac{n+3}{2}\right)$ sta fra due termini $a_s, a_t \left(s, t \leq \frac{n-1}{2}\right)$ mentre $a_{\frac{n+3}{2}}$ funge da b_1 o b_n e quindi figura solo in $|b_1, b_2|$ oppure in $|b_{n-1}, b_n|$ quindi vi saranno $\left(n - \frac{n+3}{2} + 1\right) 2 - 1 = n - 2$ (3) intervalli $|b_h, b_{h+1}|$ contenenti $\left(a_{\frac{n+1}{2}}, a_{\frac{n+3}{2}}\right)$.

---

(1) Vedi la prima nota relativa alla dimostrazione del lemma 7.

(2) Se $i < n - i$; $i \leq n - i - 1$; $i \leq \frac{n-1}{2}$.

Se $i > n - i$; $i \geq n - i + 1$; $i + 2 \geq n - i + 3$; $i + 1 \geq \frac{n+3}{2}$; ma siccome $i \neq \frac{n+1}{2}$ si avrà $i + 1 \geq \frac{n+5}{2}$; $i \geq \frac{n+3}{2}$.

(3) Da notare che qui $i > n - i$.

Inoltre sempre per la III e la IV è $\mu_{ij} = i + n - j - |i - n + j|$ se $i \neq n - j$ poichè se $i < n - j$ cioè se $i \leq \frac{n-3}{2}$ [1] ogni termine $a_r\,(r \leq i)$ sta fra due termini $a_s\,, a_t\,(s\,, t \geq j+1)$ in quanto se vi fosse qualche a_s od $a_t\,(s\,, t < j+1)$ allora si troverebbero degli intervalli $|\,b_h\,, b_{h+1}\,|$ con $\frac{h+h+1}{2} \leq \frac{i+n-i-1}{2} < \frac{n}{2}$ contro le note proprietà di simmetria delle coppie di termini $b_h\,, b_{h+1}$. Quindi sarà $\mu_{ij} = 2\,i$ se $i < n - j$. Con analogo ragionamento applicato agli $a_r\,(r \geq j+1)$ si dimostra che $\mu_{ij} = 2\,(n-j)$ se $i > n - j$ [1] ed allora sarà per $i \neq n - j$

$$\mu_{ij} = min\,[2\,i\,,\, 2\,(n-j)] = i + n - j - |\,i - n + j\,|$$

Infine se $i = n - j$ cioè se $i \leq \frac{n-1}{2}$ [1], per la III e la IV si ha $\mu_{ij} = 2\,i - 1 = 2\,(n-j) - 1$. Infatti gli i termini $a_r\,(r \leq i)$ stanno ciascuno fra 2 termini $a_s\,,\, a_t$; essendo questi ultimi in numero di $i+1$ e non potendo essere $s\,, t < j$ per ragioni di simmetria (altrimenti si avrebbero dei $|\,b_h\,, b_{h+1}\,|$ con $\frac{h+h+1}{2} \leq \frac{i+n-i-1}{2} < \frac{n}{2}$) gli $i+1$ termini $a_s\,,\, a_t$ dovranno essere $a_j\,,\, a_{j+1}\,,\, a_{j+2}\,, \ldots\, a_n$ di cui non va considerato a_j in quanto interno ad $(a_r\,,\, a_{j+1})$; ora, poichè solo a_i tra gli a_r con $r \leq i$ starà, per le note proprietà si simmetria, fra a_j ed a_s con $s \geq j+1 = n-i+1$, avremo un intervallo di meno fra gli intervalli $|\,b_h\,, b_{h+1}\,|$ che contengono $(a_i\,,\, a_{i+1})$ in confronto al caso di $i \neq n - j$; cioè sarà :

$$\mu_{ij} = \mu_{i,n-i} = 2\,i - 1 = 2\,(n-j) - 1$$

---

[1] Se $i < n - j$; $i \leq n - j - 1$; $i < n - i - 1$; $i < \frac{n-1}{2}$; $i \leq \frac{n-3}{2}$.

Se $i = n - j$; $i \leq n - i - 1$; $i \leq \frac{n-1}{2}$.

Se $i > n - j$; $i \geq n - j + 1$; $j > n - j + 1$; $j \geq n - j + 2$; e, poichè n è dispari, $j \geq \frac{n+3}{2}$; $j + 1 \geq \frac{n+5}{2}$.

La dimostrazione si ripete analogamente per le V e le VI tenendo presente di ragionare per λ_i, prima con $i \neq \frac{n-1}{2}$, poi per $i = \frac{n-1}{2}$, per μ_{ij} con $i \neq n-j$ e poi per μ_{ij} con $i = n-j$ ricordando in quest'ultimo caso di partire dagli $i = n-j$ termini a_r $(r \geq j+1)$ che stanno ciascuno fra due termini a_s, a_t che per ragioni di simmetria non dovranno avere $s, t > i+1$ poichè altrimenti si avrebbe qualche $|b_h, b_{h+1}|$ con

$$\frac{h+h+1}{2} \geq \frac{i+2+j+1}{2} = \frac{n-j+2+j+1}{2} > \frac{n+2}{2}.$$

Rimane così completamente dimostrato il teorema.

* * *

Giova osservare che dai lemmi 7 e 13 e da un nostro precedente lavoro « *Del massimo valore dell'indice di oscillazione in una successione di termini al variare in tutti i modi possibili l'ordine di questi* » Metron Vol. XVII N. 3-4, discende che le successioni massimanti I, II e III, IV, V, VI e le loro inverse (cioè leggendole da destra a sinistra) appartengono al gruppo delle successioni massimanti l'indice semplice di oscillazione:

$$O = \frac{1}{n-1} \sum_{i=1}^{n-1} |b_i - b_{i+1}|$$

Dal punto di vista del calcolo pratico del massimo dell'indice quadratico di oscillazione sopra determinato basterà disporre i termini della successione effettiva nell'ordine indicato dalle I o II se n è pari e dalle III o IV o V o VI se n è dispari, indi si calcolerà l'indice stesso.

Calcoliamo, a titolo di applicazione e di comparazione con quello semplice, ${}^2O$, $max\ {}^2O$, $\frac{{}^2O}{max\ {}^2O}$ sulle due seguenti serie storiche su cui sono già stati calcolati [1] O, $max\ O$, $\frac{O}{max\ O}$.

---

[1] C. Gini, *Corso di Statistica*, a cura di S. Gatti e C. Benedetti 1952-1953, Veschi, Roma. Pagg. 250-52.

CAMBIO DEL DOLLARO

	Teorico *	Ufficiale		Teorico *	Ufficiale
1913	5,24	5,25	32	19,96	19,47
14	5,15	5,23	33	17,87	15,53
15	6,70	6,06	34	15,38	11,68
16	7,90	6,56	35	15,84	12,13
17	8,54	7,54	36	17,56	14,22
18	11,50	7,86	37	19,18	19,00
19	11,87	8,80	38	22,53	19,00
20	13,99	21,13	39	23,95	19,23
21	20,28	23,63	40	27,43	19,80
22	20,61	21,21	41	27,54	19,32
23	20,00	21,83	42	27,35	19,01
24	20,38	22,99	43	39,31	19,00
25	21,65	25,18	44	146,10	66,25
26	22,87	25,93	45	344,65	100,00
27	20,14	19,61	46	421,65	219,62
28	19,22	19,02	47	600,31	298,17
29	18,60	19,09	48	585,05	574,67
30	18,36	19,09	49	590,63	589,76
31	18,96	19,17	1950	538,40	624,78

* Ottenuto moltiplicando il rapporto tra l'indice dei prezzi all'ingrosso in Italia (base 1913) e l'indice corrispondente per gli U.S.A. per la parità della lira col dollaro nel 1913 assunta pari a 5,24 lire per dollaro.

Naturalmente per la stessa permutazione della (1) sarà $^2O \geq O$ poichè si tratta rispettivamente di media quadratica e media semplice delle stesse $n - 1$ differenze; quindi, poichè, come è già stato fatto notare sopra, le distribuzioni massimanti l'indice quadratico di oscillazione massimizzano pure quello semplice si avrà $max\ ^2O \geq max\ O$.

Calcolato ^2O sulle due serie originarie, per calcolare $max\ ^2O$ basterà graduare prima le due serie in ordine non decrescente di grandezza dei termini e quindi ordinarle secondo la (II) del

lemma 7 (poichè nel nostro caso n è pari ed $n/2$ è dispari) indi calcolare gli indici quadratici. Confrontando questi valori con i risultati corrispondenti ottenuti con gli indici semplici avremo :

$$^2O = 49{,}84\,;\ max\ ^2O = 295{,}27\,;\ \frac{^2O}{max\ ^2O} = 0{,}1688$$

$$O = 18{,}60\,;\ max\ O = 175{,}62\,;\ \frac{O}{max\ O} = 0{,}1059$$

serie cambio teorico

$$^2O = 52{,}51\,;\ max\ ^2O = 253{,}70\,;\ \frac{^2O}{max\ ^2O} = 0{,}2070$$

$$O = 17{,}13\,;\ max\ O = 134{,}02\,;\ \frac{O}{max\ O} = 0{,}1278$$

serie cambio ufficiale

Si nota che per gli indici assoluti quelli quadratici sono maggiori per la serie del cambio ufficiale, al contrario di quanto avviene per i corrispondenti indici semplici ; per i massimi invece si hanno valori maggiori per il cambio teorico sia con gli indici quadratici che con quelli semplici e per gli indici relativi si hanno valori maggiori per il cambio ufficiale sia con gli indici quadratici che con quelli semplici.

SUMMARY

In this work the author demonstrates the following theorem on the mean square successive difference (1) :

THEOREM — *Let* $(a_i) = a_1 \leq a_2 \leq \ldots \leq a_n$ *be a non decreasing sequence of positive real numbers not all equal. The maximum value*

(1) The mean square successive difference of a sequence of real numbers : $b_1, b_2, \ldots b_n$ is the non-negative square root :

$$^2O = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (b_i - b_{i+1})^2}\,.$$

that ^2O *may reach in correspondence with the* $n!$ *permutations of* (a_i) *is :*

$$max\ ^2O = \sqrt{\frac{\sum_{i=1}^{n-1} \lambda_i (a_i - a_{i+1})^2 + 2 \sum_{i<j} \mu_{ij} (a_i - a_{i+1}) (a_j - a_{j+1})}{n - 1}}$$

where :

$$\text{for } n \text{ even} \begin{cases} \lambda_i = n - |2i - n| \quad \text{for } i \neq \frac{n}{2}; \quad \lambda_{\frac{n}{2}} = n - 1 \\ \mu_{ij} = i + n - j - |i - n + j| \quad \text{for } i \neq n - j; \\ \mu_{ij} = 2i - 1 = 2(n - j) - 1 \quad \text{for } i = n - j \end{cases}$$

$$\text{for } n \text{ odd} \begin{cases} \lambda_i = n - |2i - n| \quad \text{for } i \neq \frac{n-1}{2}, \frac{n+1}{2} \\ \begin{cases} \lambda_{\frac{n-1}{2}} = n - 1; \ \lambda_{\frac{n+1}{2}} = n - 2 \ \text{if} \ a_{\frac{n+1}{2}} \geq \frac{1}{2}\left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right) \\ \text{or} \\ \lambda_{\frac{n-1}{2}} = n - 2; \ \lambda_{\frac{n+1}{2}} = n - 1 \ \text{if} \ a_{\frac{n+1}{2}} \leq \frac{1}{2}\left(a_{\frac{n-1}{2}} + a_{\frac{n+3}{2}}\right) \end{cases} \\ \mu_{ij} = i + n - j - |i - n + j| \quad \text{for } i \neq n - j; \\ \mu_{ij} = 2i - 1 = 2(n - j) - 1 \quad \text{for } i = n - j. \end{cases}$$

From the proof of this theorem the maximizing and minimizing sequences are determined.

The author moreover makes a comparison betwen the above maximum and the analogous one of the *mean absolute successive difference* that he has determined in a note issued in No. 3-4 Vol. XVII of « Metron ».

At last some applications to time series have been made.

STEFANIA GATTI

Sul massimo di un indice di anormalità

TEOREMA — *Dati n reali* $a_1, a_2, \ldots a_n$ *non tutti uguali, il massimo valore del rapporto* $\frac{2\,\sigma^2}{{}^1S^2}$ [1] *dove*

$$\sigma^2 = \frac{\sum_{i=1}^{n} (a_i - M_1)^2}{n}, \qquad {}^1S = \frac{\sum_{i=1}^{n} |\, a_i - M_1 \,|}{n}, \qquad M_1 = \frac{\sum_{i=1}^{n} a_i}{n},$$

al variare in tutti i modi possibili le intensità delle a_i $(i = 1, 2, \ldots n)$ [*purchè non siano mai tutte uguali*] *è uguale a* :

$$max\, \frac{2\,\sigma^2}{{}^1S^2} = n$$

Per dimostrare il precedente teorema occorre premettere i seguenti due lemmi.

LEMMA 1° — *Dati n reali* $a_1, a_2, \ldots a_n$ $(a_i \geq 0\,;\ i = 1, 2, \ldots n)$ *il massimo valore che può assumere l'espressione* $n\,M_2^2 = \sum_{i=1}^{n} a_i^2$, *al variare in tutti i modi possibili le intensità dei singoli termini* a_i $(i = 1, 2, \ldots n)$ *compatibilmente con le condizioni* :

---

[1] Il rapporto $\frac{2\,\sigma_2}{{}^1S^2}$ è un noto indice statistico che serve per misurare l'iperbinomialità e l'ipobinomialità delle distribuzioni (cfr., ad esempio, C. GINI, *Asimmetria e anormalità delle distribuzioni statistiche* in « Metron », Vol. XVI, N. 1-2, 1951.

a) *invarianza della media aritmetica* $M_1 = \frac{\sum_{i=1}^{n} a_i}{n}$

b) *permanenza della limitazione* $a_i \geq 0$

è uguale a :

$$max\ n\ M_2^2 = n^2\ M_1^2$$

DIMOSTRAZIONE — Se $M_1 = 0$ ciò significa che le a_i debbono essere tutte nulle e il lemma è ovviamente dimostrato, se $M_1 \neq 0$ si vede facilmente che il massimo valore di $\Sigma\, a_i^2$ compatibilmente con le condizioni a) e b) si ha quando un termine è diverso da zero e i rimanenti $n - 1$ termini assumono il valore zero. Si abbiano infatti le seguenti distribuzioni :

$$0, 0, 0, \ldots 0, k \qquad (k = \Sigma\, a_i) \tag{1}$$

$$b_1, b_2, b_3, \ldots b_{n-1}, b_n \qquad (\Sigma\, b_i = \Sigma\, a_i;\ b_i \geq 0) \tag{2}$$

dove $k = \Sigma\, a_i$ perchè siano rispettate le condizioni a) e b) e le b_i $(i = 1, 2, \ldots n)$ sono reali qualsiasi soddisfacenti alle condizioni a) e b) e tra le quali vi sono almeno due termini diversi da zero. Siano inoltre $\overline{M}_2$ e $\overline{\overline{M}}_2$ le medie quadratiche rispettivamente delle distribuzioni 1) e 2). Sarà :

$$n\ \overline{M}_2^2 = k^2 = (\Sigma\, a_i)^2 = (\Sigma\, b_i)^2 = \Sigma\, b_i^2 + \sum_{i \neq j} b_i\, b_j =$$

$$= n\ \overline{\overline{M}}_2^2 + \sum_{i \neq j} b_i\, b_j > n\ \overline{\overline{M}}_2^2$$

Il massimo cercato è quindi dato da :

$$max\ n\ M_2^2 = n\ \overline{M}_2^2 = n^2\ M_1^2$$

LEMMA 2° — *Dati n reali qualsiasi* $a_1 \leq a_2 \leq \ldots \leq a_n$ *il massimo valore dell'espressione* $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (a_i - M_1)^2}{n}}$ *al variare in tutti*

i modi possibili le intensità dei singoli termini $a_i\,(i = 1, 2, \ldots n)$ *compatibilmente con la condizione*:

$$a)\ {}^1S = \frac{\sum_{i=1}^{n} |\, a_i - M_1 \,|}{n} = \text{costante}$$

è dato da:

$$max\ \sigma = \sqrt{\frac{n}{2}}\, {}^1S$$

Dimostrazione — La media aritmetica M_1 delle a_i sia un valore compreso tra a_r e a_{r+1} $(a_r \leq M_1 \leq a_{r+1})$. Lo scostamento semplice medio ${}^1S$ si potrà allora scrivere:

$$n\ {}^1S = \sum_{i=1}^{r} (M_1 - a_i) + \sum_{i=r+1}^{r} (a_i - M_1)$$

Facciamo ora variare le intensità delle a_i in tutti i modi possibili ma in modo che M_1 sia costante e compresa tra *l'*r^{mo} e l'$(r+1)^{mo}$ termine, ossia in modo che ci siano sempre r termini non superiori a M_1 e $n - r$ termini non inferiori a M_1.

Ricordando che la somma degli scarti positivi dalla media aritmetica è uguale alla somma dei valori assoluti degli scarti negativi, per calcolare il massimo valore di σ sotto le condizioni ora poste e quella che ${}^1S$ sia costante, basterà calcolare:

I) il massimo di $\sum_{i=1}^{r} (M_1 - a_i)^2$ facendo variare gli r termini $(M_1 - a_i) \geq 0$ compatibilmente con le condizioni:

a) $\sum_{i=1}^{r} (M_1 - a_i) = \text{costante} = \frac{n\ {}^1S}{2}$

b) permanenza della limitazione $(M_1 - a_i) \geq 0$

II) il massimo di $\sum_{i=r+1}^{n} (a_i - M_1)$ facendo variare gli $n - r$ termini $(a_i - M_1) \geq 0$ compatibilmente con le condizioni:

a) $\sum_{i=r+1}^{n} (a_i - M_1) = \text{costante} = \frac{n\ {}^1S}{2}$

b) permanenza della limitazione $(a_i - M_1) \geq 0$

Per il lemma 1° il primo massimo si avrà per le seguenti distribuzioni delle $(M_1 - a_i)$ $[i = 1, 2, \ldots r]$:

$$0, 0, 0, \ldots 0, \quad \frac{n\,{}^1S}{2} \qquad \text{se} \quad r \neq 1$$

$$\frac{n\,{}^1S}{2} \qquad \text{se} \quad r = 1$$

ossia per le seguenti distribuzioni delle a_i $(i = 1, 2, \ldots r)$

$$M_1, M_1, \ldots M_1, \quad \left(M_1 - \frac{n^1S}{2}\right) \qquad \text{se} \quad r \neq 1$$

$$\left(M_1 - \frac{n\,{}^1S}{2}\right) \qquad \text{se} \quad r = 1$$

e il secondo massimo si avrà per le seguenti distribuzioni delle $(a_i - M_1)$ $[i = r + 1, r + 2, \ldots n]$:

$$0, 0, 0, \ldots 0, \quad \frac{n\,{}^1S}{2} \qquad \text{se} \quad r \neq 1$$

$$\frac{n\,{}^1S}{2} \qquad \text{se} \quad r = 1$$

ossia per la seguente distribuzione delle a_i $(i = r + 1, r + 2, \ldots n)$:

$$M_1, M_1, \ldots M_1, \quad \left(M_1 + \frac{n\,{}^1S}{2}\right) \qquad \text{se} \quad r \neq 1$$

$$\left(M_1 + \frac{n\,{}^1S}{2}\right) \qquad \text{se} \quad r = 1$$

Si avrà cioè, comunque sia r:

$$max \sum_{i=1}^{r} (M_1 - a_i)^2 = \frac{n^2\,{}^1S^2}{4}$$

$$max \sum_{i=r+1}^{n} (a_i - M_1)^2 = \frac{n^2\,{}^1S^2}{4}$$

da cui si ricava :

$$max\, n\sigma^2 = max \sum_{i=1}^{r} (M_1 - a_i)^2 + max \sum_{i=r+1}^{n} (a_i - M_1)^2 = \frac{n^2\, {}^1S^2}{2}$$

da cui si ha infine :

$$max\, \sigma = \sqrt{\frac{n}{2}}\, {}^1S \qquad (3)$$

Osserviamo ora che l'espressione del massimo trovato non dipende nè dal valore di M_1 nè dal posto che M_1 occupa tra i termini della distribuzione. Ad uguale risultato si perviene per M_1 diverso e compreso tra due termini qualsiasi della distribuzione. Ne segue che la (3) rappresenta il massimo valore che può raggiungere lo scostamento quadratico medio σ compatibilmente con l'unica condizione che sia costante lo scostamento semplice medio ${}^1S$.

Le distribuzioni massimanti sono, come si può vedere facilmente, le infinite distribuzioni in cui vi sono : *un termine uguale a* $k - \frac{n\, {}^1S}{2}$, *un termine uguale a* $k + \frac{n\, {}^1S}{2}$ *e i rimanenti* $n - 2$ *termini uguali a* k dove k è un reale qualsiasi.

Da quanto precede si può quindi concludere che, data una distribuzione di n reali $a_1, a_2, \ldots a_n$, tra scostamento quadratico e scostamento semplice medio, sussiste sempre, comunque siano le a_i, la relazione :

$$\sigma \leq \sqrt{\frac{n}{2}}\, {}^1S$$

* * *

Passiamo ora a dimostrare il nostro teorema.

Calcoliamo in un primo tempo il massimo valore del rapporto $\frac{2\,\sigma^2}{{}^1S^2}$ al variare delle a_i compatibilmente con la condizione ${}^1S =$ = costante.

Indichiamo con ${}^1\overline{S}$ il valore costante $\neq 0$ dello scostamento semplice medio. Il massimo cercato si otterrà ovviamente facendo

variare le a_i in maniera che σ sia massimo compatibilmente con la condizione ${}^1S =$ costante $= {}^1\overline{S}$.

Per il lemma 2^0 detto valore massimo di σ è:

$$max\ \sigma = \sqrt{\frac{n}{2}}\ {}^1\overline{S}$$

e quindi si ha:

$$max\ \frac{2\,\sigma^2}{{}^1\overline{S}^2} = 2 \cdot \frac{n\,{}^1\overline{S}^2}{2} \cdot \frac{1}{{}^1\overline{S}^2} = n$$

Il valore massimo trovato è cioè indipendente dal valore ${}^1\overline{S}$ e quindi ad uguale risultato si perviene facendo assumere ad ${}^1S$ qualsiasi valore (tranne ovviamente il valore 0, nel qual caso il rapporto assume la forma indeterminata $\frac{0}{0}$). Ne segue che:

$$max\ \frac{2\,\sigma^2}{{}^1S^2} = n \qquad (4)$$

Si vede inoltre facilmente da quanto esposto che le distribuzioni massimanti sono le infinite distribuzioni in cui vi sono: *un termine uguale a* $k - h$, *un termine uguale a* $k + h$ *e i rimanenti* $n - 2$ *termini uguali a* k dove k è un reale qualsiasi ed h è un reale maggiore di zero.

Se si pone come condizione che le a_i varino restando costante la media aritmetica M_1 si vede subito che il massimo del rapporto $\frac{2\,\sigma^2}{{}^1S^2}$ sarà ancora uguale ad n, ma le distribuzioni massimanti saranno del tipo: *un termine uguale a* $M_1 - h$, *un termine uguale a* $M_1 + h$ *e i rimanenti* $n-2$ *termini uguali a* M_1 con $h > 0$.

Se si pone come condizione che le a_i varino entro l'intervallo $\alpha \mid\!-\!\mid \beta$ il massimo sarà ancora n e le distribuzioni massimanti saranno del tipo: *un termine uguale a* k, *un termine uguale a* h *e i rimanenti* $n - 2$ *termini uguali a* $\frac{k+h}{2}$ dove k e h sono due qualsiasi valori compresi nell'intervallo $\alpha \mid\!-\!\mid \beta$.

Se si pone come condizione che sia costante la media aritmetica M_1 e le a_i varino entro l'intervallo $\alpha \vdash\!\dashv \beta$ il massimo sarà ancora uguale ad n e le distribuzioni massimanti saranno del tipo: *un termine uguale a* $M_1 - h$, *un termine uguale a* $M_1 + h$ *e i rimanenti* $n - 2$ *termini uguali a* M_1 dove h è un reale maggiore di zero che non può superare il più piccolo dei due valori $|M_1 - \alpha|$ e $|M_1 - \beta|$.

Ricordando che, per un noto teorema, è

$$\frac{\Sigma a_i}{n} \leq \sqrt{\frac{\Sigma a_i^2}{n}} \quad \text{se } a_i \geq 0 \ (^1)$$

e tenendo presente la (4) si potrà infine osservare che: *dati* n *reali non tutti nulli* $x_1, x_2, \ldots x_n$ *tali che* $\Sigma x_i = 0$ *è sempre*:

$$\frac{1}{\sqrt{n}} \leq \frac{\sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}}{|x_1| + |x_2| + \ldots + |x_n|} \leq \frac{1}{\sqrt{2}} \tag{5}$$

Quanto sopra, geometricamente significa che: *dato lo spazio ad* n *dimensioni* $(x_1, x_2, \ldots x_n)$ *se si unisce con l'origine degli assi un punto qualsiasi dell'iperpiano* $x_1 + x_2 + \ldots + x_n = 0$, *tranne naturalmente l'origine ,si ottiene un segmento la cui lunghezza è maggiore o uguale a* $\frac{1}{\sqrt{n}}$ *della somma delle lunghezze delle proiezioni di tale segmento sugli assi e minore o uguale a* $\frac{1}{\sqrt{2}}$ *della somma delle lunghezze di dette proiezioni, il primo segno di uguaglianza della* (5) *verificandosi soltanto per quei punti dell'iperpiano* $\Sigma x_i = 0$ *o le cui coordinate sono tutte uguali tra loro in valore assoluto e il secondo segno di uguaglianza verificandosi soltanto per tutti i punti dell'iperpiano per cui due sole coordinate sono diverse da zero, ossia per tutti i punti di intersezione dell'iperpiano con i piani coordinati e soltanto per essi.* In questo ultimo caso, come si può vedere facilmente, il nostro teorema si riduce al noto teorema: *la lunghezza della diagonale di un quadrato è uguale alla lunghezza del lato moltiplicata per* $\sqrt{2}$.

---

(1) Veggasi, ad esempio, C. GINI, *Corso di Statistica* a cura di S. GATTI e C. BENEDETTI (nuova edizione aggiornata) a. a. 1954-55, Ed. Veschi, Roma, 1955.

SUMMARY

The authoress demonstrates the following theorem:

THEOREM — *Let* $a_1, a_2, \ldots a_n$ *be a sequence of real numbers; varying in all possible ways the single intensities of* a_i *under the condition:*

$$\sum_{i=1}^{n} | a_i - M_1 | > 0$$

we have:

$$max \frac{2\sigma^2}{{}^1S^2} = n$$

where:

$$M_1 = \frac{1}{n} \sum_{i=1}^{n} a_i ; \; {}^1S = \frac{1}{n} \sum_{i=1}^{n} | a_i - M_1 | ; \; \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i - M_1)^2 .$$

MARY JEAN BOWMAN

The analysis of inequality patterns : a methodological contribution

The measurement of inequality has been a subject of extensive study for many years. The aim, usually, has been to procure a single coefficient that may be used as a summary measure of " inequality " for the distribution as a whole. Important and well-known contributions to the arsenal of social scientists working with such material are the Paretos α, Gini's δ and concentration ratio, and Gibrat's coefficient of concentration. The Gini concentration ratio has one advantage over others : its meaning is independent of the degree of fit of the empirical distribution to any particular *a priori* mathematical formula.

The inconsistencies and ambiguities that may arise from the use of different summary indexes of inequality were analyzed by the writer some years ago. (1) No summary measure is free of the condition that in its computation some particular method of weighting differences or deviations is involved. Because of this restriction, contrasts in underlying socio-economic structures are sometimes concealed when reliance is placed upon comparisons by means of a single summary index. In the present paper, as in the earlier one, emphasis is placed upon the representation of *entire* distributions. The present discussion takes one additional step, and focuses upon what, for lack of a better term,

(1) *A Graphical Analysis of Personal Income Distribution in the United States*, « American Economic Review », XXXV, No. 4 (1945), 607-28. Reprinted in *Readings in the Theory of Income Distribution*, Philadelphia : Blakiston., 1946.

may be designated as the " structuring of inequalities " *within* distributions. It is the comparison of total patterns, rather than of single indexes of inequality that is here the primary concern.

That the patterning as well as the over-all degree of inequality is an economically and socially relevant attribute of frequency distribution would seem to be axiomatic. Yet it has received comparatively little attention. The two main exceptions that have come to the attention of the author are R.R. Schutz's analysis of the slopes of the Lorenz curve and Ernesto P. Billeter's study of variations in the value of Gini's δ as applied to different segments of a given distribution. (2)

The simplest graphical technique available for study of inequality patterns is of course the Lorenz curve, when analyzed in terms of its *shape* and not merely in terms of the area between that curve and the " line of equality. " Schutz took the further step of plotting the slopes of the Lorenz curve, which are in fact simply the ratios to the arithmetic mean at the various percentiles of the distribution. (3) Billeter's study is much more elaborate and extremely suggestive in many ways. However, the values of δ in the various sectors of a distribution are not independent of one another. This is a defect that cannot be ignored in using his method for an analysis of the " structure of inequality. "

Whatever the advantages or defects of each of these approaches, a special drawback lies in the fact that for many problems the available data lack the information on aggregates or averages

---

(2) R. R. Schutz, *On the Measurement of Income Inequality*, « American Economic Review », XLI, No. 1 (1951), 107-22. For a discussion of his article see G. Rosenbloth's note in the « American Economic Review », XLI, No. 3 (1951) 135-37. Ernesto P. Billeter, *Über die Messung der Einkommens-Konzentration*, Bern, Francke, 1949.

(3) These ratios form a J-shaped curve that is a rather blunt instrument for comparison of " inequality structures " in various distributions. A related possibility is to plot the logarithms of the ratios against income percentiles, the latter on a probability scale. While this facilitates visual comparisons, interpretation becomes more difficult. There is little to justify focussing on the logarithms of ratios to the arithmetic mean.

necessary for the construction of Lorenz curves or the computation of Gini's δ.

There are, of course, other very useful and simple statistical measures, such as the relative quartile deviation or ratios of quartiles or deciles to the median and to each other. But building up a picture of inequality structure from such measures becomes something of a patchwork operation. Hence it has seemed worthwhile to search for other possibilities.

The substantive materials analyzed in the present methodological explorations are the 1949 distributions of gross farm incomes in five American states : North Carolina, South Carolina, Mississippi, Georgia, and Iowa. Study of a number of socio-economic attributes of farmers in these states incited the writer to a renewed interest in the whole question of the patterning of inequalities within distributions. Examination of miscellaneous sets of data concerning farms in these states suggested the presence of distinct and consistent state configurations that were repeated from one variable to another. (4) It is the author's intention subsequently to apply the techniques developed in this paper to socio-economic distributions of several kinds and in different countries.

In simplest terms the general method consisted in the graphing of small-segment dispersions against the percentiles of the distributions, using dispersion measures associated with curve types that provide good initial approximations to the data. (The curve types used had in common the characteristic of not requiring information concerning aggregate values of the variable in question).

Whatever the basic curve type or " reference formula " used, the associated inequality charts provide simultaneously : (1) a picture of the structuring of inequalities within each distribution, (2) a comparison of absolute degrees of inequality at corresponding percentile points in the several empirical distributions, and (3) evidence concerning the nature of the deviations of the empirical distributions from the selected formula. The third of these features is especially important in that it offers useful hints

(4) For recent statistical evidence see C. Arnold Anderson, *Economic Inequalities in Southern Agriculture*, « Rural Sociology », March 1954.

Percentages of Farms with gross incomes exceeding designated values.

Race, Farm Type and State	Gross Farm Income					
	$250	$400	$600	$1000	$1200	$1500
ALL RACES :						
All Farms						
N.C.	76.9	71.6	65.9	56.7	52.2	46.8
S.C.	75.2	66.9	57.6	43.3	37.9	31.2
Ga.	75.5	68.2	60.0	46.5	41.2	35.0
Miss.	77.1	66.4	55.1	36.3	29.8	22.4
Iowa	96.1	94.6	92.6	89.5	88.1	86.1
	$2500	$4000	$5000	$6000	$10000	$25000
N.C.	28.5	12.1	6.8	4.4	1.1	0.1
S.C.	16.5	6.7	4.3	3.2	1.5	0.4
Ga.	20.3	10.5	7.4	5.8	2.6	0.5
Miss.	9.4	4.1	2.8	2.3	1.3	0.4
Iowa	78.7	66.9	58.4	50.5	25.4	3.6
	$250	$1200	$2500	$5000	$10000	$25000
ALL RACES :						
Commercial Farms						
N.C.	100.0	78.2	42.3	10.3	1.6	0.2
S.C.	100.0	62.4	26.8	7.0	2.4	0.6
Ga.	100.0	66.8	32.4	11.9	4.0	0.7
Miss.	100.0	47.9	14.7	4.6	2.0	0.7
Iowa	100.0	95.3	85.0	63.2	27.1	3.9
WHITE :						
All Farms						
N.C.	*	48.1	26.6	7.3	1.3	0.2
S.C.	*	37.9	19.5	6.7	2.6	0.7
Ga.	*	41.7	22.6	9.3	3.3	0.6
Miss.	*	28.8	11.8	5.0	2.3	0.8
Commercial Farms						
N.C.	100.0	77.9	43.1	11.8	2.1	0.3
S.C.	100.0	70.8	36.4	12.4	4.7	1.2
Ga.	100.0	71.5	38.7	16.0	5.7	1.0
Miss	100.0	53.1	21.8	9.3	4.3	1.5
NEGRO :						
All Farms						
N.C.	*	66.0	33.7	5.5	0.2	0.04
S.C.	*	37.6	12.0	1.2	0.1	—
Ga.	*	39.9	12.5	1.7	0.1	0.04
Miss.	*	30.8	6.3	0.5	0.1	0.02
Commercial Farms						
N.C.	100.0	78.8	40.3	6.7	0.3	0.05
S.C.	100.0	54.1	17.2	1.7	0.1	—
Ga.	100.0	55.5	17.4	2.4	0.2	0.06
Miss.	100.0	43.7	9.0	0.8	0.1	0.03

* Data on non-commercial farms lack the details necessary to distinguish those with incomes be ow $250 from those between $250 and $1200.

Sources: All farms, all races - except at $1200, $5000 and $25000, from 1950 Census of Agriculture, Volume II, p. 766, Table 5; all others computed from Economic Area Table 12 and State Table 16.

as to the nature of the socio-economic factors that may be at work behind the empirical data. By revealing contrasts in the patterning of dispersion through the ranges of the empirical distributions to be compared (i.e., contrasts in the " structuring of inequalities "), such charts may both suggest new hypotheses and provide checks upon previous hypotheses relating to underlying factors.

The initial problem in applying this method is the identification of a curve type that will adequately describe some central position among the empirical distributions to be compared. Then by a suitable transformation the formula may be expressed in such a manner that cumulative frequencies take on a linear form, with the slope of the associated straight line constituting a summary measure of the degree of " equality " or " inequality. " A perfect fit of the empirical data to the formula would of course involve a constant slope in the *actual* curve ; plotting that slope against the cumulative percentiles would give a horizontal line. Systematic deviations of the empirical data from the formula chosen would be reflected in systematic variations of slopes through the range of the distribution ; these variations of slope would show up (Chart 2) as deviations of the empirical " patterns of inequality " from the inequality patterns implicit in the underlying formula.

Five initial types of formulae or " reference " curves were examined. After discarding one of these, (5) sets of associated

---

(5) The fifth type was the family of compound binomial distributions. *A priori* a reasonable hypothesis in the interpretation of highly skewed distributions is provided by the probability rationale of compound binomials. These were discussed in considerable detail by C. C. Boissevain some years ago : *Distribution of Abilities Depending upon Two or More Independent Factors*, « Metron » XIII (1935), 49-49. H. T. Davis' work along these lines is of course well known, and his earlier discussions were drawn upon by Boissevain.

For the present study, the fit of the empirical gross farm income distributions to compound binomial formulae was tested graphically, assuming three and then four independent factors and adjusting the origin and dispersion to secure the best approximations. It was evident that a five-factor compound binomial would be less adequate than the four-factor case which in turn was better than the three-factor form. However, plotted

or derivative inequality distributions were constructed for the remaining four. These four curve types are as follows, taking $F(x)$ as the proportion of cases with gross incomes equal to or exceeding x, and x_0 as the median gross income.

$$F(x) = \frac{1}{2}\left(\frac{x}{x_0}\right)^{\alpha_1} \qquad x \gtrless x_0 \qquad \alpha_1 < 0 \tag{1}$$

$$G = \frac{F(x)}{1 - F(x)} = \left(\frac{x}{x_0}\right)^{\alpha_2}; \qquad F(x) = \frac{x^{\alpha_2}}{x^{\alpha_2} + x_0^{\alpha_2}} \tag{2}$$

$$x \gtrless 0 \qquad \alpha_2 < 0$$

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{\omega}^{\infty} e^{-\frac{\omega^2}{2}}\, d\omega \tag{3}$$

$$\textit{where} \quad \omega = a \log x + b$$

$$F(x) = 1 - \frac{2}{\pi} tan^{-1}\left(\frac{x}{x_0}\right)^{\alpha_3} \qquad \alpha_3 > 0 \tag{4}$$

Note that in all four cases $F(x) = 0$ for $x = \infty$, and $F(x) = \frac{1}{2}$ for $x = x_0$. With the exception of formula (1), $F(x) = 1$ when $x = 0$. (In formula (1) any application below $F(x) = \frac{1}{2}$ would clearly be inappropriate).

---

cumulatively on double log paper the points of maximum curvature in the empirical distributions were in every case sharper than the " best " model and the upper halves of the empirical distributions did not have the concavity that characterized the model. The compound binomial hypothesis failed to provide the reference type required on three counts: First, use of an arbitrary negative origin was necessary to secure the best fit, and for gross income data this was unreasonable *a priori* and hence incompatible with any *a priori* rationale that might be argued in defense of the model. Second, simplicity in both computations and interpretations would be forfeited without sufficient justification in the empirical results. And third, the income data all deviated from the compound binomial curves in the same directions in the various percentile ranges of the distributions.

The first of these is the Pareto formula, applied to the upper half of the distribution. The second is a similar formula for $\frac{F(x)}{1 - F(x)}$. The third is an expression of the logarithmic transform of the normal curve in terms of $F(x)$. To conserve space the fourth formula has been omitted from the charts presented in this article. It has the advantage over a Pareto-type of analysis in that splicing at the median is not required. However, above the median the pattern shown is intermediate between the Pareto-type charts and the other two. Since *a priori* there seems to be no special reason for considering this formula on grounds of any interpretative value, it will not be discussed any further.

Inequality Patterns in Gross Farm Income: All Races Together

The curves shown on the three panels of Chart 1 are applications of the first three formulae to the gross farm income data of the top sections of Table 1. In each case, as previously noted, a perfect fit to the formula would be reflected in a straight line of constant slope. In Chart 2 the reciprocals of the slopes of the Chart 1 curves are plotted against the income percentiles (changing the sign in the first two cases); these reciprocals show the patterning of inequalities through the range of the empirical distributions for which data are available. In each case, the greater the values of Y, the greater the inequalities as measured by the particular index involved.

The method used in estimating the values for Chart 2 is very simple. ([6]) First, the known points on a frequency distribution are expressed in units suitable to the linear transformations required in applying the particular formula under examination. Let any three consecutive points so expressed be $x_{-1}\,y_{-1}$, $x_0\,y_0$, $x_{+1}\,y_{+1}$. The slope at $x_0\,y_0$ is then equal to

$$\frac{y_0 - y_{-1}}{x_0 - x_{-1}} + \frac{y_{+1} - y_0}{x_{+1} - x_0} - \frac{y_{+1} - y_{-1}}{x_{+1} - x_{-1}}.$$

([6]) The author is indebted to Fil. Lic. GUNNAR K. O. KULLDORFF, lecturer in statistics at the University of Lund, Sweden, for suggesting this method and for other helpful criticisms.

It should be noted that x_0, does not here refer to the median income, but merely to the income (or rather, in all these cases, the logarithm of the income) at any point at which a measure of slope is being made. For the first panel, based on the Pareto formula, the y's are simply the logarithms of $F(x)$. For the second panel they are the logarithms of $\frac{F(x)}{1 - F(x)}$, and for the third panel, based on the log-normal formula, they are the number of standard deviation units on a normal distribution encompassed between $F(x)$ and the median.

For Iowa this method provides estimates only up to the top quartile. Hence the percentile at $24,000 was estimated by fitting a 2nd degree equation to the top three points in the Iowa data (expressed in units appropriate to each formula). Then the slope at the $24,000 point was estimated as in other cases. A similar method was used to extend Negro farm income "inequality curves" down to lower percentile levels than could otherwise have been included (see the next section and Chart 3).

While it is widely recognized that for most income distributions the Pareto formula fits at best only in the upper part of the income scale, there is still sufficiently strong attachment to this formula in many quarters to require that it be considered here. If it fits reasonably well over most of the distribution above the median, the lower half of the distribution might be related to the inverse of that formula. However, it is evident from Chart 1A that all five distributions have a marked concavity below the top quintile (or decile) level. In the upper ranges the Mississippi and South Carolina curves are clearly convex and the North Carolina curve is slightly so. The upper part of the Georgia curve is almost straight and that for Iowa continues slightly concave.

The similarities and contrasts among the states in the nature of their deviations from the Pareto formula are demonstrated in Chart 2 A, where $Y_1 = -1/a$. (The Y values are plotted on a logarithmic scale to facilitate comparison (7). Relative

(7) It should be noted that Y_1 measures the elasticity of the number of income recipients with a given income or more, with respect to income. When this elasticity is low, Y_1 is high.

to that formula, dispersion is in every case greatest at the lowest income levels (in this application of the Pareto formula, at the median) ; dispersion is at a minimum around the top decile or higher. At the extreme top, dispersion relative to the Pareto formula increases markedly for the Mississippi and South Carolina distributions, tends to stabilize in North Carolina and Georgia, and continues to decline in Iowa. In view of the poor fits of the Pareto formula between the median and the top decile, a check on the fit of the inverse of that formula below the median, where the data are less adequate (Iowa excepted) is omitted.

With the exception of the very top of the distributions, above the 95th percentile, it is clear that the other two formulae tested in Chart 2 perform much better than the Pareto model. The log-normal model provides the best approximation through the range from the median to the 95th percentile, though all the curves except that for Iowa turn up markedly at the top as compared with this formula. The second formula is in general intermediate between the first and third. While some people might consider formula (2) the most satisfactory as a compromise, there are two reasons for discarding it. (1) In all of the actual series the inequality curves derived from this formula have a decided negative slope throughout the range from the median to the top decile. (2) Like formula (4), formula (2) offers little promise for interpretation in terms of socio-economic postulates.

That the logarithmic transform of the normal curve yields a distinct curve type unlike any of the Pearsonian type distributions has been demonstrated by others. (8) Its rationale in probability theory was discussed by Gibrat over twenty years

---

(8) For this demonstration see Edwin B. Wilson and Jane Worcester, *The Logarithmic Transform*, « Review of Economic Statistics », XXVII(1945), 17-22. R. Gibrat argued ably and at length for the applicability of the log-normal curve in simple or modified form to a wide range of socio-economic variables, testing his " law of proportional effect " on distributions of income, wealth, inheritance, size of enterprise (as measured by numbers of employees), and size of towns : *Les Inegalites Economiques*, Thesis for the doctorate, University of Lyon, Paris, 1931. However, it should be pointed out that like most students of inequality measurement,

ago, and important additions to and modifications of that analysis were contributed by Kalecki in 1945 ([10]). More rigorous mathematical considerations aside, such a distribution has a certain *a priori* plausibility. However, in the last analysis the validity of any hypothesis depends upon its correspondence with the facts. Does the normal distribution of log x meet the requirements of a reference model for use in the present context ?

Comparison with the normal distribution is facilitated by plotting $F(x)$ on a probability scale, ([11]) against log x as the abscissa. (Parenthetically it should be noted that this choice of ordinates has an advantage over the more usual double logarighmic curves in that the lower as well as the upper end of the distribution is " stretched out ", contributing to a visual appreciation of differences among the curves throughout the entire range. ([12])) A log-normal distribution would appear on this diagram as a straight line, and the steeper that line the less the standard deviation of log x. A symmetrical distribution of log x with a kurtosis in excess of the normal pattern would be reflected in a curve on this chart that was symmetrical about the median, but with a steeper slope in the central range than at the ex-

---

Gibrat was interested in a *single summary* measure of inequality. By contrast, the concern of the present study is with the patterning of inequalities reflecting systematic deviations from a master model or curve type. (For a further comment on Gibrat, see footnote [13]).

([9]) *Op. cit.*

([10]) M. Kalecki, *On the Gibrat Distribution*, « Econometrica XIII (1945), 161. Kalecki improves upon Gibrat's analysis of the rationale in probability teory for his " law of proportional effect " as applied to the simple log-normal distribution and he adds a theoreticalexplanation of the modified case in which there is a normal distribution of log $(x-A)$.

([11]) On probability paper the scale of $F(x)$ becomes the w of Formula (3).

([12]) J. and M. I. Dufrenoy argued that this kind of curve should be used in the analysis of net income distributions. While they seem, mistakenly, to use it as a rationalization of the Paretian pattern, they nevertheless introduce some interesting hypotheses concerning potential affects of public policy and economic developments upon income distributions. *La distribution des biens et la distribution des aptitudes*, « Journal de la Societe de Statistique de Paris », LXXXIX (1948), 321-33. See also, W. C. Helme, *The Relation Between Rents and Incomes, and the Distribution of Rental Values*, « Bell System Technical Journal », I (1922), 82-109.

tremes. An arithmetically normal distribution would start comparatively flat and become continuously steeper throughout. Georgia comes the closest to a log-normal distribution, while Mississippi departs from the normal in one direction, Iowa and for the most part North Carolina in the other direction. However, even the Iowa distribution is much closer to a log-normal than to an arithmetically normal form.

In Chart IIC, $Y_3 = 1/a$, where a is the slope of w on log x (see formula (3)). That is, the vertical scale on this diagram is the standard deviation of log x implied by the slope of the corresponding Chart 1C curve at the percentile point under consideration. If the empirical data fitted the log-normal formula perfectly, Y_3 would of course be constant. ([13]) A symmetrical distribution of log x but with a high kurtosis would yield on this chart a symmetrical U-shaped curve. A distribution that was arithmetically normal would appear as a curve dropping steeply from infinity at the low income margin, then declining more gradually through the center range, and finally approaching zero at the top of the income scale.

([13]) While a better approximation to a straight line on Chart 2 C could be obtained for Iowa, North Carolina, and Mississippi if we used log $(x\text{-}A)$ on the abscissa, with the values for A (positive or negative) chosen to secure the best fit in each case, there is no *a priori* justification for assuming values of A other than zero in analyzing these gross income data. By introducing the extra constant the comparability of both summary inequality measures such as GIBRAT's C and the inequality patterns of a chart analogous to Chart 2 C would be at least partially destroyed, and their socio-economic meaning would be blurred. This is a major weakness of GIBRAT's analysis, for the value of A affects the value of his inequality measure C. GIBRAT was driven to use this modification of the simpler formula because of his desire to secure a summary inequality measure; in terms of the analysis presented in this paper, he was seeking to get a fit such that the value of Y would be constant. Where the patterning of inequality rather than a summary measure is the focus of attention, this dilemma is avoided. (It is of course quite legitimate to use a normal distribution of log $(x\text{-}A)$ when the purpose is simply to obtain as good a fit to a particular distribution as possible. This kind of adjustment is discussed by WILSON and WORCESTER, *op. cit.* They present a method of determining the most "likely" value of A for any given empirical distribution).

Only in the case of Iowa (where comparatively few farm incomes are in the lowest census income category) do the data permit analysis of the lower percentiles of the distribution. However, within the ranges for which data are adequate, two facts are evident. (1) There are in each case marked deviations from the inequality pattern implied by the reference formula, but there is a greater range of comparative constancy than in the other panels. (2) There are clear contrasts among the states in the patterns revealed.

As compared with a normal distribution of log x, the lower half of the Iowa farm population is spread out increasingly the lower the percentile position on the income scale; the upper half of the Iowa distribution closely approximates the formula pattern. In North Carolina inequalities diminish continuously as compared to the log-normal formula from at least the lowest third to the top fifth of the distribution, though there is a distinct rise again at the extreme top. Mississippi inequalities in the lower and median ranges are less than in any of the other southern states and up to the top decile they more nearly fit the log-normal model, but the Mississippi and South Carolina distributions turn upward sharply above the top decile of the population. This suggests a more hierarchical socio-economic structure in South Carolina and Mississippi with a small privileged minority clearly distinguished from the bulk of the farm population. This trait of these two states is indicated by other well-documented socio-economic attributes of the farm populations of the various states. However, the South Carolina pattern as a whole is quite different from that in Mississippi so far as relative spread through the middle ranges is concerned; the gap between the top five percent and the top decile of the South Carolina farmers appears in an accentuated form because of the previous decline of the curve from the relatively high degree of spread (i.e., relative inequality) in the vicinity of the median.

The fourth panel of Chart 2 duplicates the third panel except that it applies to the distribution of "commercial farms" only. (The census definition of commercial farms excluded part-time farms and all farms, regardless of other considerations, with gross incomes from sales below $250.) While exclusion of the

part-time farms makes for greater comparability among those remaining, the $250 limit involves an arbitrary distortion-especially important in Mississippi where many farm families had total dollar incomes below that level. As would be expected, inequality among low income farms in Iowa and among low and middle income farms in the southern states is considerably less when "non-commercial" farms are excluded. The effects on the North Carolina distribution are most striking. The removal of the non-commercial farms lowers the percentile position of the upward turning point in the South Carolina and Mississippi curves. The data are less complete for commercial farms alone; hence the curves drawn on the last panel of Chart 2 are considerably less realiable in detail than those of the third panel. [14]

Within-Race Patterns of Farm Income Inequalities

In view of the comparatively satisfactory fit of the empirical distributions of all farm incomes in the various states to the logarithmic transform, this curve type was tested on the two racial groups within the southern farm population. Unfortunately, the census data by race in the all-farm series provide adequate information only for the top third or fourth of the distributions except for North Carolina. Census tabulations for the commercial farms alone permit analysis from the median upward among whites and for approximately the top third of the Negro farmers.

---

[14] In the all-farm series census data were available for computation of the GINI's δ for only the upper 10 to 20 percent of the southern farm distributions. However, for commercial farms (all races) there was sufficient information to delineate the patterns from the median upward. The results, which should be compared with panel 4 of Chart 2, were as follows: All five curve drop considerably (by about a third or more) from the median to the 65th percentile. Beyond this the South Carolina curve becomes horizontal, while the Georgia curve continues to drop throughout. The Mississippi curve approximates a horizontal configuration above the 80th percentile, and the North Carolina curve turns up slightly. (The special computation necessary to extend the Iowa curve above the top quartile was not undertaken.) The contrasts between inter-distribution relations based on this measure and those shown by the other measures reflect the fact that the values of δ at the various percentile points of a distribution are not independent of one another.

A preliminary glance at Chart 3 suggests that the log-normal model is in general intermediate between the white and the Negro distributions, the former showing a greater skewness and the latter a lesser skewness than would characterize a log-normal distribution. The declining values of Y_3 that characterized the southern all-farm, all-race distributions (except Mississippi) between the median and the top decile are largely eliminated or even reversed when only whites are considered, whether for all farms or for commercial farms only. (The Mississippi white distribution for the all-farm series closely approximates the Pareto formula over the top fourth of the income range, but the Pareto distribution gives a poor fit below the 90th percentile in all the other cases.) Among Negroes, in both the all-farm and the commercial farm distributions, the general pattern seems to be one of lesser inequality than that prevailing among whites, and of less skewness than in the log-normal pattern. Partial exceptions to the latter generalization are the Mississippi distributions, which show a rather definite upturn of Y_3 in the top 5 percent of the population. It is obvious from Chart 3 that the log-normal model provides a better approximation to the Negro distributions than would any of the other formulae considered.

Conclusions

Meaningful study of inequalities requires that their patterning and not merely over-all summary measures of degree of inequality be given careful attention. To this end, a graphical analysis of the "structuring of inequalities" has been suggested. The basic method is (1) to select some linear function of a curve type that describes a central pattern relative to the empirical distributions to be compared, (2) to plot the empirical curves using this function, and then (3) to derive the slopes of the (2) curves (or their reciprocals) by percentiles of the distribution.

For the gross farm income data used here the logarithmic transform of the normal curve provided the best reference model, and there is evidence that this formula may be useful in analyzing a wide variety of socio-economic data.

The basic argument of this paper, however, does not depend upon the particular formula used. Whatever the basic formula best suited to the distributions to be compared, the charts of inequality patterns provide a convenient summarization, highlighting important features of the distributions that are too commonly ignored. This approach is therefore submitted as a promising addition to the tools designed for the analysis of inequalities.

Chart 1. Cumulative Distributions; Gross Farm Incomes, 1949

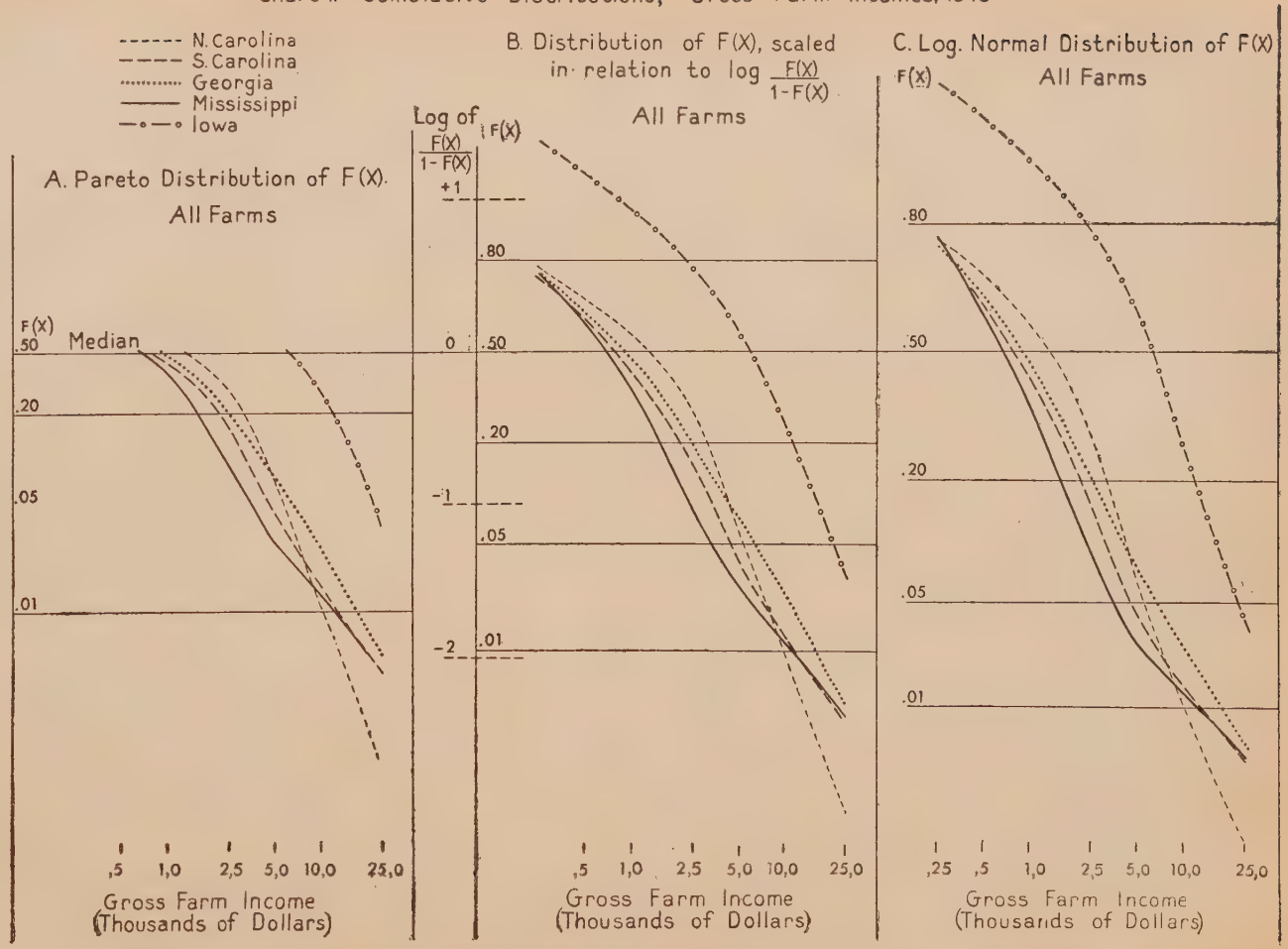


Chart 2. Inequality Patterns; Gross Farm Incomes

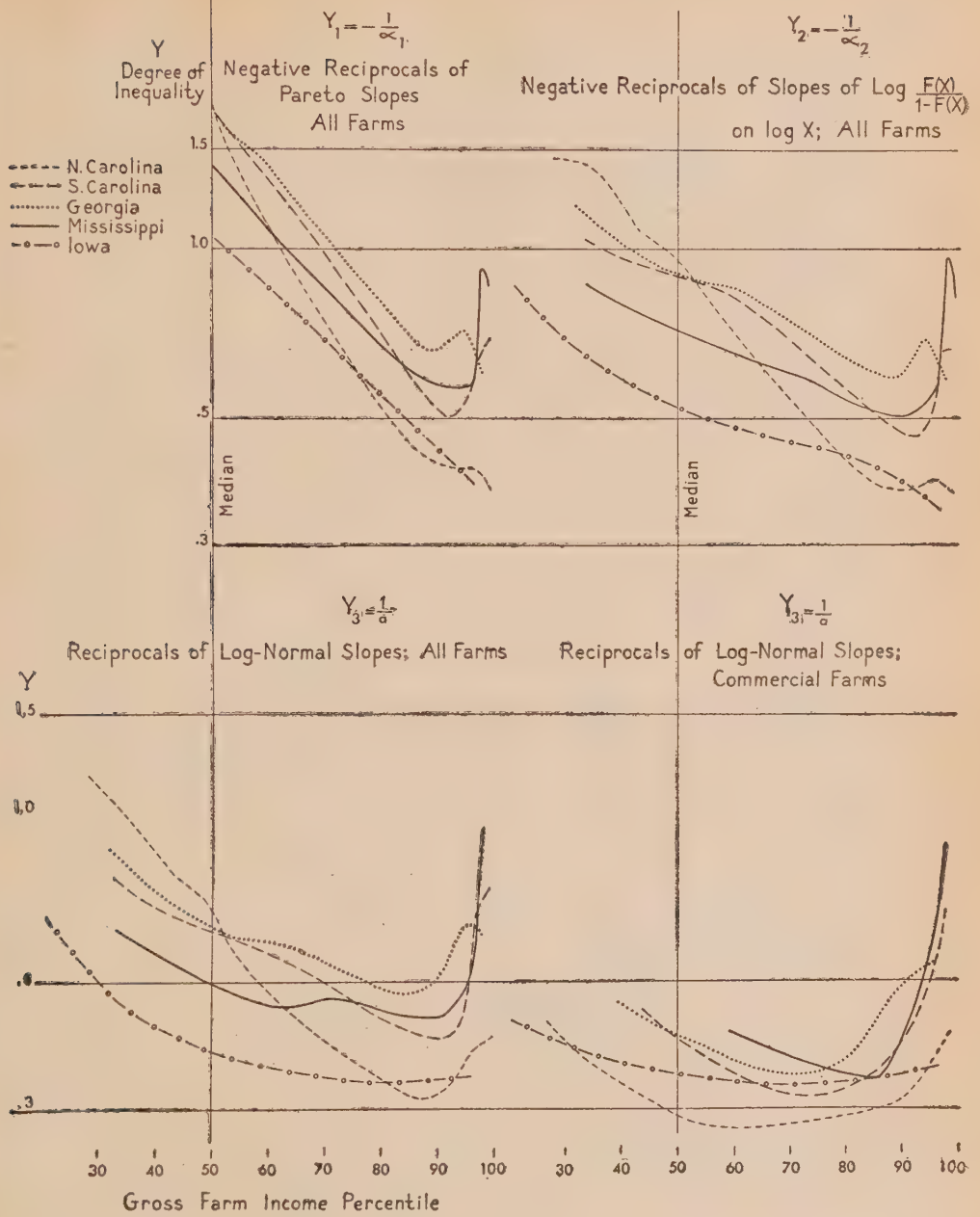
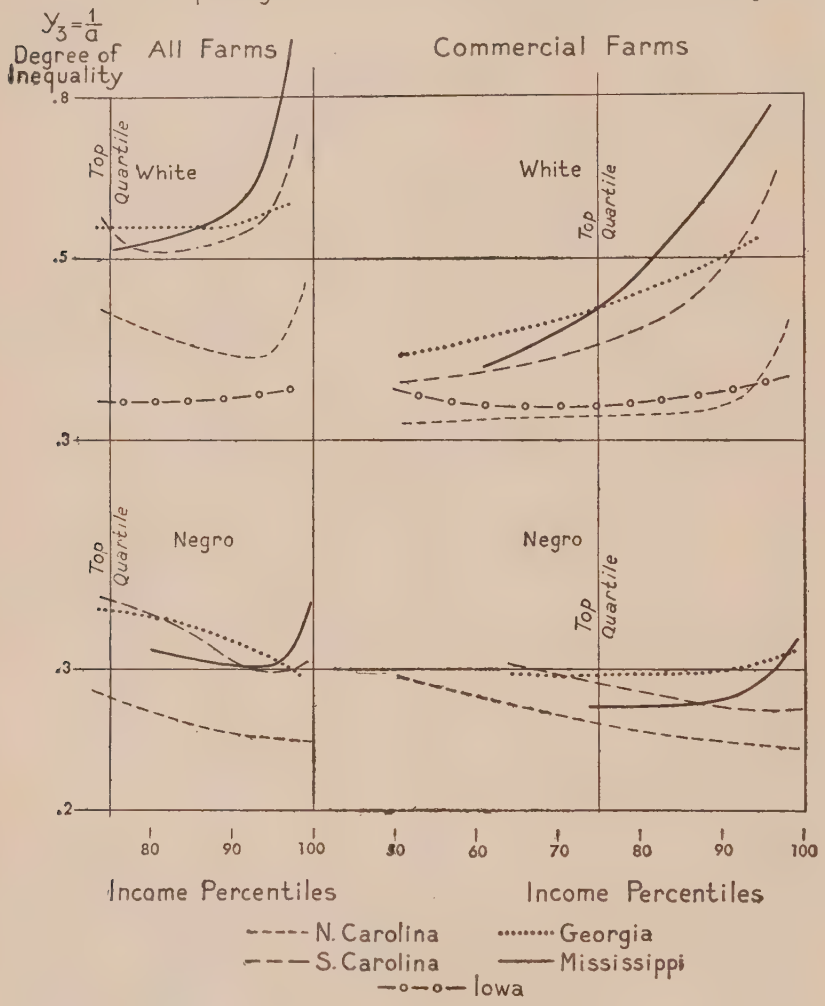


Chart 3. Inequality Patterns; Gross Farm Incomes, by Race



ELIO CARANTI

Su un procedimento approssimato per la determinazione del numero medio dei figli per matrimonio (1)

1. — Molteplici sono, com'è noto, i procedimenti escogitati per misurare la fecondità matrimoniale. Basati su materiale statistico di vario tipo, essi forniscono indici atti a rappresentare il fenomeno, nei suoi diversi aspetti, in maniera più o meno soddisfacente dal punto di vista teorico.

Per avere dati continuativi circa il numero medio dei figli per matrimonio, tali da rispecchiare le condizioni attuali della fecondità di una popolazione, si è fatto ricorso — in mancanza delle statistiche di base necessarie per l'applicazione di formule esatte — a diversi metodi approssimati.

Il più semplice, che chiameremo metodo A), consiste nel rapportare il numero dei nati legittimi in un anno di calendario, N_x, al numero dei matrimoni, M_x, celebrati nell'anno stesso, calcolando cioè

$$F'_x = \frac{N_x}{M_x} \qquad (1)$$

Esso è corretto, come si vedrà più oltre, limitatamente al caso in cui il numero dei matrimoni resti costante nel tempo, in quanto solo una piccola frazione dei nati proviene dai matrimoni conclusi nello stesso anno di calendario.

Dall'osservazione che il maggior contributo relativo alle nascite avvenute in un dato anno è normalmente fornito dai matri-

(1) Questo lavoro è stato oggetto di una comunicazione al Seminario di Statistica della Facoltà di Scienze Statistiche, Dem. ed Attuariali della Università di Roma il 24 Giugno 1954.

moni dell'anno precedente, M_{x-1}, deriva il metodo B), espresso da

$$F''_x = \frac{N_x}{M_{x-1}} \qquad (2)$$

In base all'ipotesi di variazione lineare del numero annuale dei matrimoni è stato proposto il metodo C), nel quale si considerano al denominatore della frazione i matrimoni, M_{x-t}, celebrati in un anno anteriore a quello cui si riferiscono le nascite di un periodo uguale all'intervallo medio t fra data del matrimonio e nascita di un figlio di ordine di generazione qualsiasi; si ha quindi:

$$F'''_x = \frac{N_x}{M_{x-t}} \qquad (3)$$

2. — Il difetto comune a questi procedimenti approssimati risiede nel fatto di tener conto esclusivamente dei matrimoni celebrati in un solo anno di calendario, cosicchè — a prescindere dal verificarsi delle ipotesi di partenza — i risultati risentono delle non indifferenti oscillazioni che la nuzialità presenta nel tempo, anche a breve periodo.

Per ovviare a tale inconveniente, circa venticinque anni fa il Gini propose ed applicò un metodo più elaborato (1), fondato su un principio al quale aveva già fatto ricorso al fine di misurare la mortalità infantile (2).

Seguendo tale metodo, che chiameremo D), il numero dei nati legittimi in un certo anno di calendario viene rapportato ad un *numero virtuale di matrimoni*, $\overline{M}_x$, determinato facendo una media aritmetica ponderata dei matrimoni — celebrati nello

---

(1) C. GINI, *Di un procedimento per la determinazione del numero medio dei figli per matrimonio*. Relazione presentata al Congresso Int. per gli Studi sulla Popolazione (Roma, 1931) e riprodotto in *Saggi di Demografia*, Roma, 1934, pp. 13-39; vedi anche « Metron », vol. X nn. 1-2, 1932 e *Sur une mèthode pour déterminer le nombre moyen des enfants légitimes par mariage*, « Rev. Inst. Int. de Stat. », 1933 n. 1.

(2) C. GINI, *Sulla mortalità infantile durante la guerra*, « Atti della Soc. It. di Ostetr. e Ginec. », vol. XIX, 1919, riprodotto in *Problemi Sociologici della Guerra*, Bologna, 1921 e, in lingua inglese, in « Eugenics Review », gennaio 1920; vedi pure *On a new method for calculating the infantile death rate according to the month of death*, « Rev. Inst. Int. de Stat. », 1934 n. 3.

stesso anno e negli anni precedenti — da cui quelle nascite derivano.

La ponderazione è effettuata in base alla frequenza con cui i matrimoni celebrati negli anni da $x - s$ ad x (s indica la distanza dei matrimoni più lontani nel tempo) danno luogo a nascite nell'anno x. Per la pratica applicazione, il metodo può essere rappresentato sotto la forma

$$F_x^{IV} = \frac{N_x}{\overline{M}_x} = \frac{N_x}{\sum_{i=0}^{s} M_{x-i} \cdot P_{x-i}} \tag{4}$$

essendo

$$P_{x-i} = \frac{\frac{n_{x(x-i)}}{M_{x-i}}}{\sum_{i=0}^{s} \frac{n_{x(x-i)}}{M_{x-i}}} \tag{5}$$

i « fattori di ponderazione », dove $n_{x(x-i)}$ indica il numero dei nati nell'anno x provenienti dai matrimoni celebrati nell'anno $x - i \left(\sum_{i=0}^{s} n_{x(x-i)} = N_x \right)$ [3].

Sostituendo nella (4) ai P_{x-i} i rispettivi valori dati dalla (5) si ottiene la seguente espressione

$$F_x = \sum_{i=0}^{s} \frac{n_{x(x-i)}}{M_{x-i}} \tag{6}$$

la quale costituisce la formula esatta per il calcolo del numero medio dei figli legittimi per matrimonio. Nell'ipotesi che il numero dei matrimoni sia stato costante nel tempo, ad esempio uguale a M_x, nell'anno x di calendario avremmo infatti non N_x, ma

$$\overline{N}_x = \sum_{i=0}^{s} n_{x(x-i)} \cdot \frac{M_x}{M_{x-i}} \tag{7}$$

---

[3] La (5) è stata data dal Gini sotto la forma

$$P_{x-i} = \frac{n_{x(x-i)} \cdot \frac{M}{M_{x-i}}}{\sum_{i=0}^{s} n_{x(x-i)} \cdot \frac{M}{M_{x-i}}} \tag{5'}$$

dove M rappresenta un numero costante, qualunque, di matrimoni.

nascite, che possiamo chiamare *numero virtuale delle nascite.* Dividendo per M_x si ottiene la (6) (4).

3. — Il metodo D) posa quindi su basi assolutamente rigorose.

Notevole è d'altronde la sua utilità pratica: per pochissimi paesi, infatti, sono disponibili dati continuativi sulle nascite legittime classificate secondo l'anno di celebrazione del matrimonio, necessari per il calcolo della (6), mentre per l'applicazione della (4) basta determinare — mediante una indagine eseguita *una tantum* — i valori dei P_{x-i} dati dalla (5). Questi permettono di calcolare, conoscendo il solo numero annuale delle nascite complessive e dei matrimoni, prima $\overline{M}_x$ e poi F^{IV}_x (5).

La superiorità del metodo D) rispetto ai metodi A), B) e C) — che pure possono applicarsi avendo a disposizione gli stessi

---

(4) Non mi consta che, al momento in cui scriveva il Gini, fosse da considerarsi pacifica la validità della (6), che dal procedimento approssimato proposto logicamente discende. Riferendosi all'analogo caso della mortalità infantile, anzi, L. Galvani (*Calcolo delle probabilità di morte in generale, e applicazione alla misura della mortalità infantile*, « Annali di Statistica » serie VI vol. XVII, 1931, pp. 10-13) aveva giudicato a priori non giustificabile la scomposizione del tasso di mortalità in tassi parziali, relativi a distinti contingenti di nati.

Ancora di recente, poi, sono state riprese — sempre a questo riguardo — formule che danno al problema soluzione diversa, sia pure, a mio avviso, errata. Vedi in proposito E. Caranti, *Su alcuni metodi di misura della mortalità infantile*, « Atti della XI e XII Riun. Scient. della Soc. It. di Stat. », Roma, 1954, pp. 365-370 e *Metodi di misura della mortalità infantile*, comunicazione presentata alla 28° Sessione dell'Ist. Int. di Stat., Roma, 6-12 settembre 1953.

(5) Questo vantaggio sembra essere sfuggito al Mortara (vedi l'articolo *Sui metodi per lo studio della fecondità dei matrimoni*, « Giornale degli Econ. e Riv. di Stat. », vol. XLIII n. 12, 1933, pp. 893 e segg.) il quale ha criticato il procedimento come se la (4) fosse sostitutiva della (6), che è invece ovviamente preferibile applicare disponendo dei dati necessari.

In particolare il metodo *D*) presenterebbe « lo svantaggio di non condurre direttamente alla determinazione dei tassi di fecondità del tipo $\frac{n_{x(x-)i}}{M_{x-i}}$ che sono proprio i risultati più importanti ricavati da queste indagini, e l'altro svantaggio di introdurre il concetto, perfettamente inutile, di *numero virtuale dei matrimoni.* » *(segue)*

elementi — era prevedibile a priori. È però sembrato opportuno procedere ad una verifica pratica, anche per valutare la portata delle riserve formulate (6) circa la possibilità di mantenere immutati i P_{x-i} per lunghi periodi, o quanto meno in periodi caratterizzati da notevoli modificazioni della fecondità.

A tal fine ci si è giovati del fatto che dal 1930 al 1950 l'Istituto Centrale di Statistica italiano ha pubblicato le cifre delle nascite secondo l'anno di celebrazione del matrimonio. Ciò ha permesso di determinare il numero medio dei figli per matrimonio mediante la (6), assumendo i risultati come dati di riferimento per controllare l'approssimazione dei vari metodi presi in considerazione.

Per quanto riguarda il metodo D), si sono estesi i calcoli effettuati dal Gini per il trentennio 1903-32 al successivo ventennio 1933-52, impiegando gli stessi valori dei P_{x-i}, derivanti da un'apposita indagine compiuta sulle nascite dal 1927 (7).

In effetti si potrebbe anche pensare di operare, al fine di un calcolo approssimato di F_x, sul numeratore anzichè sul denominatore, scomponendo cioè, mediante adeguati « fattori di separazione » da tenere fissi, il numero complessivo delle nascite N_x nei suoi elementi componenti $n_{x(x-i)}$. Ma se si vuole eliminare — come si dovrebbe, per garantire la bontà dei risultati — l'influenza del diverso numero di matrimoni da cui i singoli contingenti di nati derivano, si ricade inevitabilmente nel metodo *D*). Ciò è stato laboriosamente dimostrato da Ptoukha (*Mèthodes pour calculer les taux de mortalité infantile selon les mois de l'année*, « Rev. Inst. Int. de Stat. », 1934 n. 2), ma appare chiaramente anche dalla (7).

Il Livi (*Trattato di demografia. Le leggi naturali della popolazione*, Padova, 1940, pp. 117-118) accoglie invece sostanzialmente il concetto di numero virtuale dei matrimoni, interpretandolo come una media armonica degli M_{x-i}, ponderati in base agli $n_{x(x-i)}$ e dando di conseguenza una dimostrazione della (6) diversa da quella indicata nel presente lavoro. L'A. esprime però riserve sulla (4), dichiarando che, per applicare « coefficienti di correzione normalizzati e fissi », occorrerebbe ponderare la media armonica degli M_{x-i}, anzichè con gli $n_{x(x-i)}$, con i rapporti $\frac{n_{x(x-i)}}{N_x}$. Tale osservazione non sembra giustificata ed è d'altra parte in contrasto con quanto il Livi stesso ammette nella dimostrazione da lui precedentemente data della (6), in cui esegue precisamente la ponderazione degli M_{x-i} in base agli $n_{x(x-i)}$.

(6) Vedi L. Livi, op. cit., p. 118, nota (1).

(7) Agli allievi della Facoltà di Scienze Statistiche, Demografiche ed Attuariali Enrico Lattanzi e Virginia Piacentini sono dovuti rispettiva-

4. — I dati delle tabelle 1 e 2 dimostrano in modo evidente la bontà delle approssimazioni fornite dal metodo D), sia in via assoluta che in relazione agli altri metodi approssimati più comunemente adottati.

Pur trattandosi di un periodo demograficamente molto perturbato, con notevoli oscillazioni della nuzialità e della fecondità, si rilevano infatti divergenze massime, rispetto ai risultati ottenuti applicando la (6), del 2,1 % nel 1945 e del 2,2 % nel 1950, mentre nella maggior parte dei casi (e precisamente per 15 anni su 20) le differenze sono inferiori all'1 %. La media è di 0,80 %, a fronte di 12,53 % per il metodo A), che presenta un massimo del 38,7 % nel 1943; di 9,9 % per il metodo B), con massimo di 34,0 % nel 1944; di 12,45 % per il metodo C) con $t=6$ e di 8,50 % con $t = 8$, a cui corrispondono divergenze massime del 56,9 % nel 1950 e del 24,5 % nel 1945.

Per gli anni nei quali il metodo D) dà luogo, con i P_{x-i} adottati, alle peggiori approssimazioni, sono stati calcolati i valori effettivi degli stessi P_{x-i} (con i max $= 29$), riportati nella tab. 3. Come si vede, le modificazioni del sistema dei fattori di ponderazione sono sensibili, tanto che l'indice di dissomiglianza percentuale rispetto alla distribuzione del 1927 assume il valore di 26,9 per il 1935, di 35,4 per il 1940, di 27,9 per il 1943, di 19,5 per il 1945 e di 16,9 per il 1950.

Nonostante ciò, come si è visto, i risultati ottenuti col metodo D) presentano delle divergenze da quelli esatti di lieve entità. Si deve quindi ritenere provata la rispondenza del metodo stesso nella misura del numero medio dei figli per matrimonio e concludere che, ove non si disponga di una rilevazione continuativa che permetta il calcolo della (6), esso può essere proficuamente applicato. Si intende che l'uso di un certo sistema di fattori di ponderazione non può estendersi, per risparmiare una indagine apposita, a paesi diversi da quello per cui furono calcolati, se non dopo avere controllato la somiglianza delle caratteristiche della

---

mente i calcoli per l'applicazione della (6) e le altre elaborazioni. I risultati dei primi differiscono da quelli ottenuti e pubblicati dal Lenti (*La fecondità matrimoniale in Italia dal* 1930 *al* 1950, «Giorn. degli Econ. e Ann. di Econ.», anno XII nn. 9-10, 1953, p. 541) perchè sono stati considerati i soli nati vivi, come dal Gini, e non il totale delle nascite, come dal Lenti.

Tabella I — *Approssimazione dei vari metodi di calcolo del numero medio dei figli per matrimonio.*

(Differenze percentuali dalla (6))

ANNI	D)	A)	B)	C)	
				$t = 6$	$t = 8$
1950	+ 2.24	— 5.10	— 0.94	+56.89	+19.96
1949	+ 0.64	— 8.15	—14.18	+53.60	—23.69
1948	+ 0.47	—15.14	—25.39	+15.99	+ 6.54
1947	— 0.33	—27.93	—16.86	+18.14	+ 9.80
1946	— 0.66	—29.04	+ 2.61	— 3.83	— 0.46
1945	— 2.07	— 9.57	+29.67	—11.26	—24.47
1944	— 0.29	+33.87	+34.04	— 9.43	— 7.24
1943	— 0.92	+38.74	+ 5.95	—19.29	+ 6.63
1942	— 0.37	+ 6.78	+12.12	— 3.05	— 1.87
1941	— 0.11	+13.52	— 1.11	+ 8.01	+ 7.17
1940	+ 1.47	+ 1.02	— 1.62	+ 1.49	+18.52
1939	+ 0.57	— 3.32	— 3.99	+ 7.58	+12.99
1938	+ 0.59	— 3.44	—16.85	+17.13	+ 3.44
1937	— 0.04	—19.71	— 4.31	+ 9.72	+ 5.24
1936	+ 0.61	— 5.80	+ 3.65	— 1.66	+ 4.53
1935	+ 1.05	+ 4.15	— 4.20	+ 4.09	— 0.99
1934	+ 0.62	— 5.34	+ 2.10	+ 3.77	+ 0.01
1933	+ 0.94	— 1.33	+ 9.71	— 2.90	— 0.67
1932	+ 0.97	+10.22	+ 6.91	— 0.02	— 3.82
1931	+ 1.05	+ 8.51	— 1.21	+ 1.27	—13.72

Tabella 2 — Numero medio dei figli per matrimonio calcolato con vari metodi

(Nati vivi legittimi per 1.000 matrimoni)

ANNI	APPLICAZIONE DEI METODI					
	Esatto (form. 6)	D)	A)	B)	C)	
					$t = 6$	$t = 8$
1950	2.615	2.673	2.482	2.455	4.103	3.137
1949	2.757	2.775	2.533	2.366	4.235	2.411
1948	3.071	3.085	2.606	2.292	3.562	3.272
1947	3.185	3.174	2.295	2.648	2.762	3.497
1946	3.479	3.457	2.469	3.570	3.346	3.464
1945	2.857	2.798	2.584	3.705	2.535	2.158
1944	2.769	2.761	3.707	3.712	2.508	2.569
1943	2.927	2.900	4.061	3.101	2.362	3.121
1942	2.988	2.978	3.190	3.350	2.897	2.932
1941	2.985	2.982	3.389	2.952	3.224	3.199
1940	3.189	3.236	3.222	3.138	3.237	3.780
1939	3.220	3.222	3.114	3.092	3.465	3.639
1938	3.189	3.208	3.080	2.652	3.736	3.299
1937	3.156	3.154	2.534	3.020	3.462	3.321
1936	3.104	3.123	2.924	3.218	3.052	3.245
1935	3.188	3.221	3.320	3.054	3.318	3.156
1934	3.226	3.246	3.054	3.294	3.348	3.231
1933	3.261	3.292	3.304	3.578	3.166	3.239
1932	3.173	3.204	3.497	3.393	3.169	3.052
1931	3.248	3.282	3.525	3.209	3.290	3.803

Tabella 3 — Valori dei P_{x-i} percentuali.

X — *i*	1927 (Gini)	1950	1945	1943	1940	1935
x	4,43	4,56	4,36	4,69	4,14	4,39
x—1	13,99	16,15	18,74	14,53	15,00	15,54
x—2	8,79	9,57	8,39	8,01	8,59	8,42
x—3	7,79	8,21	7,49	7,47	8,17	8,35
x—4	7,10	7,06	6,47	6,32	7,07	7,10
x—5	6,45	6,50	5,76	5,67	6,41	6,46
x—6	5,76	6,26	5,12	5,26	5,90	5,93
x—7	5,28	5,07	4,66	5,04	5,35	5,38
x—8	4,82	4,60	4,31	4,76	4,98	4,78
x—9	4,42	4,14	4,05	4,48	4,47	4,48
x—10	4,02	3,84	3,87	4,31	4,20	3,85
x—11	3,67	3,37	3,54	4,09	3,85	3,74
x—12	3,33	2,99	3,35	3,75	3,53	9,33
x—13	3,03	2,71	3,14	3,52	3,10	3,04
x—14	2,73	2,44	2,81	3,16	2,94	2,66
x—15	2,48	2,21	2,59	2,90	2,62	2,39
x—16	2,18	1,96	2,30	2,49	2,31	2,06
x—17	1,94	1,75	2,01	2,22	1,34	1,67
x—18	1,74	1,52	1,67	1,91	1,57	1,37
x—19	1,49	1,26	1,44	1,53	1,25	1,11
x—20	1,24	1,08	1,18	1,25	0,94	1,01
x—21	0,99	0,82	0,91	0,88	0,66	0,91
x—22	0,74	0,64	0,66	0,61	0,52	0,66
x—23	0,55	0,44	0,45	0,40	0,38	0,50
x—24	0,39	0,33	0,28	0,26	0,26	0,34
x—25	0,27	0,22	0,16	0,19	0,19	0,23
x—26	0,16	0,14	0,09	0,13	0,13	0,13
x—27	0,12	0,08	0,06	0,07	0,08	0,08
x—28	0,05	0,04	0,04	0,04	0,04	0,04
x—29	0,05	0,04	0,10	0,06	0,01	0,05

fecondità matrimoniale. Analogamente, quando questa, per uno stesso paese, dovesse subire nel tempo rilevanti modificazioni, occorrerà controllare se il sistema adottato continui o meno a rispondere in maniera soddisfacente allo scopo.

5. — Non sembra inutile chiudere questa nota con qualche osservazione circa il significato della (6) — e quindi delle sue espressioni approssimate — agli effetti della misura della fecondità matrimoniale (8).

La (6) misura la fecondità presente dell'insieme di generazioni o coorti di matrimoni ancora prolifiche nel periodo in esame. È stato però osservato (9) che al denominatore della frazione figura il contingente iniziale dei matrimoni e non il numero delle coppie superstiti al processo di eliminazione per emigrazione, morte, separazione, uscita dall'età feconda ecc., intervenuto nel frattempo.

Senonchè la (6) può essere messa sotto la seguente forma

$$F_x = \sum_{i=0}^{s} \frac{n_{x(x-i)}}{M_{x(x-i)}} \cdot \frac{M_{x(x-i)}}{M_{x-i}} \qquad (6')$$

dove con $M_{x(x-i)}$ si indicano, degli M_{x-i} matrimoni celebrati nell'anno $x - i$, quelli ancora in grado di dar luogo a nascite nell'anno x.

---

(8) Vedi le considerazioni, molto generiche, formulate al riguardo, riferendosi al metodo *D*), da R. Kuczynski, *The measurement of population growth. Methods and results*, Londra, 1935, pp. 221-23.

(9) Il Lenti (op. cit., p. 531) rileva che, essendo ignoto il numero delle coppie ancora presenti, « questi quozienti di fecondità matrimoniale sono perlomeno approssimati per difetto, approssimazione che ovviamente tende ad aumentare a mano a mano che cresce la differenza fra l'anno solare di celebrazione del matrimonio e quello di nascita ».

Non è ben chiaro se l'A. intenda riferirsi ai singoli tassi parziali componenti la (6) o anche ai valori di F_x, come fa esplicitamente B. Colombo (*La recente inversione della tendenza della natalità*, Padova, 1951, p. 47). Ora, è vero che i primi sono inferiori a quelli che risulterebbero ponendo al denominatore il numero dei matrimoni sopravviventi, ma le considerazioni svolte nel testo dimostrano che lo stesso non può dirsi per F_x. In particolare è da rilevare che, volendo eliminare l'influenza — oltre che delle variazioni del numero dei matrimoni — delle variazioni della sopravvivenza matrimoniale, la (6) darebbe, ove questa fosse crescente, risultati superiori e non inferiori a quelli che si otterrebbero mantenendo invariati i valori della sopravvivenza.

I quozienti di fecondità relativi alle coppie superstiti sono quindi in realtà considerati, assumendo però una composizione dei matrimoni sopravviventi uguale a quella che si sarebbe avuta partendo da contingenti iniziali di matrimoni ugualmente numerosi, sottoposti al regime di eliminazione che tali contingenti hanno effettivamente sperimentato. Il che appare ovvio, riflettendo sul modo con cui è stata ottenuta la (6).

Tenendo presente ciò, la serie ottenuta unendo i dati del Gini e quelli nostri (vedi tabella 4) assume un valore più che indicativo delle variazioni della fecondità matrimoniale italiana, che risulta diminuita in mezzo secolo di circa il 40 % [10].

---

[10] Mentre questa nota era in corso di pubblicazione è apparso un articolo di L. Henry e R. Pressat (*Évolution de la fécondité en Italie*, « Population », vol. 10, n. 3, 1955) nel quale è stata studiata una serie simile a quella della tab. 4, ottenuta utilizzando i dati del Gini e quelli del Lenti, questi ultimi corretti per escludere i nati morti.

Tabella 4 — Numero medio dei figli legittimi per matrimonio in Italia dal 1903 al 1952.

(Nati vivi legittimi per 1.000 matrimoni)

ANNI	‰	ANNI	‰
1903	4.255	1928	3.426
1904	4.416	1929	3.325
1905	4.371	1930	3.505
1906	4.269	1931	3.248
1907	4.192	1932	3.173
1908	4.447	1933	3.261
1909	4.284	1934	3.226
1900	4.363	1935	3.188
1911	4.138	1936	3.104
1912	4.280	1937	3.156
1913	4.215	1938	3.189
1914	4.166	1939	3.220
1915	4.221	1940	3.189
1916	3.541	1941	2.985
1917	3.071	1942	2.988
1918	2.986	1943	2.927
1919	3.531	1944	2.769
1920	4.641	1945	2.857
1921	3.944	1946	3.479
1922	3.845	1947	3.185
1923	3.736	1948	3.071
1924	3.604	1949	2.757
1925	3.558	1950	2.615
1926	3.517	1951	2.522
1927	3.499	1952	2.509

Nota: I dati dal 1903 al 1930 sono quelli calcolati dal Gini applicando il metodo D), che è pure servito per calcolare, con gli stessi fattori di ponderazione, i quozienti relativi al 1951-52. Quelli dal 1931 al 1950 sono stati invece ottenuti, applicando la (6).

S. R. DAS

A Mathematical Analysis of the Phenomena of Human Twins and Higher plural births. Part III.

CONTENTS

A. Correction of Data for Prenatal Mortality.

1. *Introduction.*

In the two preceding papers (I & II), certain suggestions were made about the necessity of refining the theory as well as the data. It was stated (II-13, 14) that the formulae for multiple confinements should be modified on the consideration that the probability of a twinning scission for a zygote should decrease at each increase in the number of its scission. The purpose of the present paper is to give due consideration to this question and to effect the necessary modifications of the twin and the triplet formulae accordingly. Introduction of this refinement is, in fact, based on undisputed biological observations which have hitherto, been over-sighted by the mathematical investigators, all of whom (Dahlberg, '26, Jenkins, '27 & '29) have tacitly assumed that the probability of monozygotic twinning remains uniform for any number of scissions of the same zygote. Thus the frequency of monozygotic triplet sets has been taken as p^2, when p is that of the monozygotic twins in a population of confinements.

The other refinement which will also be done in this paper, is concerned with the early loss of the products of confinements, which are not usually recorded. It is obvious that unless reliable factual records regarding the amount of loss is available, no theoretical estimate therefor is acceptable. No general procedure for correction of the existing data is, therefore, envisaged here. It is widely accepted that the birth registration data are defective, particularly in respect of the record of very early cases of abortions, however carefully the data may be collected. One main reason is that the hold on life of the human embryo in the primary stage — before implantation, or even a week or

so after it — is precarious and many (estimates vary from 1 in 5 to 1 in 3) are lost unnoticed, as the unplanned abortions in that stage is harmless and painless as well. The dead or degenerated embryos are eliminated out along with other debris and some uterine blood (Barnett, 1950). This is a source of inaccuracy in the statistics of confinements, on which there exist no data and therefore, no method can, at present be devised to make a correction of the statistical material for this error.

In paper (II), certain quantities were calculated (Table IV) from the statistical material derived from the papers of Strandskov and his collaborators (1945 & 1946). The material was collected from the 'White' as well as the 'Colored' populations in the registration areas of U.S. during the period 1922-1936. The still-birth data included abortions also. In 1949, Strandskov and Bisaccia published an analysis of another material on still-births collected in a restricted region of the registration area. In the restricted region, it is believed that even the early cases of abortions (of course not those occuring in the first 2-3 weeks after the ovulation or fertilisation) are completely recorded. This material has been used in the present paper to rectify the previous results (for the 'white' alone) calculated from the data of the registration area, in general (Paper II, Table III & IV). In this connection, an empirical rule has been deduced, which establishes two linear relations between the mortality percentage of the single confinement on one side and that of the twin, or the triplet confinements on the other.

We shall, in a number of steps in the following sections, develop the process of mortality correction as explained above. But before proceding with them, it is thought necessary to make references to a few literatures on the sex in human births, as the analyses presented in this series of papers are directly concerned with frequency of sex, and sex-ratios. Gini (1951) in a recent review has elaborately discussed the influences other than chance in the occurrence of sex in human families, particularly, the influence of (1) heredity, (2) wishes of parents, (3) birth control and, (4) tendency for producing the same sex and later on its reversal to the other sex during the reproductive period of the parents. The factors, though of much importance in the

determination of the sequence or general distribution of sex in human families, are not likely to affect the sex composition of a big population and hence, we require no material modification of our present analyses in this light. One thing, however, suggests itself. A dizygotic twin pair born in a same-sexed family may have a greater tendency to be like-sexed, as the members of a pair of 2-egg twins are just like two sibs successively born of the same parents. The material, at present available, does not permit a closer analysis of the dizygotic twin data to test it.

Gunnar Dahlberg (1951) has introduced a new hypothesis that old eggs are fertilised by Y-sperms (i.e. male sex-chromosome bearing sperms) and that the young eggs are fertilised, with equal probability, by *X*- and *Y*-sperms. The hypothesis explains (qualitatively, of course, at present) at once the phenomena of the preponderance of male sex in human confinements and the greater mortality rate among the males than in females. The hypothesis is supported by the results of the experimental researches carried out by Hertwig (1911) on frog's eggs, Blandau and Young (1939) on old eggs of guinea pig, and Blandau and Jordan (1941) on old eggs of rats. Dahlberg has not made any suggestion as to what makes the old eggs fertilisable by the *Y*-sperms, in partial or complete exclusion of the *X*-sperms.

The old-egg hypothesis has the novelty of assigning to women an important part in controlling the determination of sex of the zygotes. Investigations of Gini (1905, 1908, 1911) and Slater (1944) also discovered that the women played a part in the determination of sequence or disposition of sex in the families. The present author likes to extend the old-egg hypothesis and assume that a zygote produced as a result of fertilisation of an young egg undergoes monozygotic scission with greater probability than one from an old egg. We shall prove in sections 6 and 7, that the older the zygote or the embryo, the less is its probability for monozygotic scission. If now the age of the ovum (after attaining maturity in the graafian follicle) is added to that of the zygote, i.e. if the age of the zygote be counted from the point of attainment of maturity of the ovum, then an explanation of Φ (the probability of monozygotic scission of a female zygote) being greater than μ, (the probability

of monozygotic scission of a male zygote) is naturally obtained. Further analyses are, of course, necessary before the idea is accepted, which, nevertheless, appears to the present writer as very plausible and simple.

In the preceding papers (II-10 & III-Table IV), it was suggested and proved by numerical values that Φ was greater than μ. Nothing was, however, said about the reason for such a differential twinning probability between the two sexes. If Dahlberg's old-egg hypothesis proves ultimately successful, its extension as suggested above may be regarded as the biological reason for the observed differential twinning probability.

2. *Empirical Relations between the mortality percentages among the single, the twin and the triplet confinements.*

The present writer takes the liberty in quoting a few statistical figures from the paper of Strandskov and Siemens (1946) which are included in Table 1. The percentages of mortality among the single, the twin and the triplet confinements in U.S. 'White' and 'Colored' populations are calculated from those figures and the results are shown in the same table.

From a scrutiny of the percentages of mortality as given in Table 1, it strikes that the twin mortality percentage is about double and the triplet mortality percentage nearly four times the mortality percentage of the single confinements. This is true in the case of the 'White' as well as the 'Coloured' Americans of the registration area.

In Table II, a comparison is made between the observed mortality percentages and those calculated for the twin and the triplet confinements by multiplying the single birth mortality percentage by 2 and 4 respectively. As additional data, the mortality percentages of the single and the multiple confinements in England and Wales (Stocks, '52) are also included in the same table, which also agree with the rule remarkably well. There is one peculiarity, however, that the calculated value of the mortality percentage in each case for the twins is lower than the observed value, while for the triplets the calculated values are higher than the observed ones without exception. In order to

fit in the observed data with the empirical rule, the latter is therefore improved as follows.

In the first instance it is found that coefficient of correlation between the single and the twin mortality percentages is almost unity (0.997) which means a perfect correlation between the two mortalities, a fact not difficult to understand. In such a case of perfect correlation, the two regression lines which are usually separate and subtend a definite angle between them, do merge into a single straight line, the equation of which represents the actual linear relation between the two variable quantities. Similar observations are true also for the coefficient of correlation (.997) between the single and the triplet mortality percentages and the regression equations between them.

The regression lines are given by (*a*) for twin-single : $y = 1.944\ x + 0.721$, $(r_{x,y} = .997)$ and (*b*) for triplet-single : $z = 3.592\ x + 0.381$, $(r_{zx} = .997)$. Here x, y and z stand for the mortality percentages in the case of single, twin and triplet confinements respectively.

The two linear equations do prove that the simple empirical rules which were enunciated previously in this connection were not far from the precise rules embodied in the equations (*a* & *b*).

Data for other populations on the mortality in the single and the multiple confinements should be analysed and then, the above equations can be firmly established — though it is believed that no major changes will be necessary.

In the present paper we shall make use of these two formulas to evaluate the mortality percentages in the twin and triplet confinements, when that for the single confinement is known. The equations furnish a useful tool for the statistical analyses of confinement data which omit to mention the mortality figures for some kinds and mention that for at least one kind of pregnancies — single, twin, or triplet. In table II, the computed values for the mortality percentages as obtained through the equations (*a*) and (*b*) are also given to show the closer agreement between these values and the observed ones than that between the latter and the values calculated from the simple empirical rule, cited in the first instance.

TABLE 1

A few data quoted from Strandskov and Siemens, 1946 and the percentages of mortality.

	'White' (frequencies)		'Colored' (frequencies)	
	Male	Female	Male	Female
Single :				
Total births.	14,249,501	13,355,589	1,803,341	1,709,494
Live-born.	13,755,800	12,985,885	1,672,457	1,613,306
Still-born	493,701	369,704	130,884	96,188
Still-born (p.c.) . . .	(3.464)	(2.768)	(7.258)	(5.627)
Twins :				
Total births.	321,217	309,243	50,675	50,225
Live-born.	298,139	290,671	43,232	44,224
Still-born	23,078	18,572	7,443	6,001
Still-born (p.c.) . . .	(7.185)	(6.006)	(14.688)	(11.948)
Triplets :				
Total births.	4,565	4,558	999	1,110
Live-born.	3,957	4,053	735	883
Still-born	608	505	264	227
Still-born (p.c.) . . .	(13.319)	(11.079)	(26.426)	(20.450)

TABLE 2

Observed and empirical prenatal mortality percentages compared.

CONFINEMENTS	U. S. White		U. S. Colored		England and Wales	
	Male	Female	Male	Female	Male	Female
Single :						
(Observed x)	3.464	2.768	7.258	5.627	2.748	2.511
Twins :						
Observed	7.185	6.006	14.688	11.948	6.504	5.473
($y = 2x$)	6.928	5.536	14.516	11.254	5.496	5.022
* Formula (a)	7.455	6.102	14.883	11.660	6.063	5.602
Triplets :						
Observed	13.319	11.079	26.426	20.450	10.185	8.547
($y = 4x$)	13.856	11.072	29.032	22.508	10.992	10.044
* Formula (b)	12.824	10.324	26.452	20.613	10.252	9.401

* Formulae : (a) $y = 1.944x + 0.721$ and (b) $z = 3.592x + 0.381$.

3. *Correction for Pre-natal Mortality applied to U.S. Data.*

Strandskov and Bisaccia (1949) did not give the mortality data separately for the single and the multiple births — males and females. All mortalities of each sex are taken together. From out of this it is possible to estimate the mortalities of the male as well as female single confinements with sufficient accuracy and next the mortality percentages of the twin and the triplet confinements can be determined with the help of the equations (*a*) and (*b*) of the preceding section.

The mortality percentage of the single confinement does not differ much from that of the total confinement as the frequencies of the various multiple confinements together form a small fraction of the total confinement frequency. Thus, if we again take the mortality data of Table I, we find that the male mortality percentage in the total male confinements is 3.544 and that for the female 2.844, whereas the mortality percentages in single male and female confinements are respectively 3.464 and 2.768. The male mortality in the single is, therefore, less than that in the total male by (0.080) and the same difference in the case of female is (0.076). In the case of the restricted region data, we first find the mortality percentage of each sex in the total confinements, which on reduction by 0.080 and 0.076 for the male and female respectively, should yield very nearly the mortality percentages in the single confinements. The author is, however, aware of the loop-hole of this procedure, but it is certain that the results obtained in the above manner shall differ but little from the actual values of those quantities.

Strandskov and Bisaccia (1949) estimated the numbers of males and females implanted as 3,962,829 and 3,761,093 respectively, of which, in all, 164,701 males and 121,367 females are lost by death in utero. Hence, the over-all prenatal mortality percentages of males and females are 4.156 and 3.227 respectively and these are for the total confinements. The corresponding figures for the single confinements will be about 4.076 and 3.151 according to what has been explained in the above paragraph. The prenatal twin and triplet mortalities for each sex are estimable from the formulas (*a*) and (*b*) in section 2. The results are shown in Table III.

TABLE III

Calculated prenatal mortality p. c. in the Restricted region of U.S. birth registration area, 1923-36.

	ALL CONFINEMENTS	SINGLE CONFINEMENTS	TWIN CONFINEMENTS	TRIPLET CONFINEMENTS
Male	4.156	4.076	8.645	15.022
Female	3.227	3.151	6.847	11.699

The results in Table III are numerically greater than the corresponding observed prenatal mortality percentages included in Table II. Assuming that there is no especial contingency causing higher mortality in the restricted region, the higher prenatal mortality percentages in the restricted region must be attributed to the fact that in this region, recording of abortions and still-births is more complete than in the registration area in general. Although it must be kept in mind that the cases of abortions taking place within 2-3 weeks after the ovulation can not be, and are not recorded even here, however complete the data may be in other respects under human control.

No extrapolatory mathematical device could be applied to the prenatal mortality data to estimate the mortality at the very prime of the zygotic, or the emboyonic life. Firstly, the prenatal mortality data, supplied by Strandskov and Bisaccia (1949) and those by the other authors (Tietze '48, Ciocco '40, Tauber '23, Auerbach '12, Wedervang '24, Schultz '18, Boldrini '36), do not agree mutually for various reasons.

Secondly, it seems that at the very prime of the embryonic life, certain unusual lethal factors work, which end many lives. They cease to act in the later part of the embryonic life, or are eliminated in this way by death.

Mathematical extrapolation implies regular variations, according to some mathematical law, of the physical conditions under which the variable to be estimated is changing. This condition is, therefore, not satisfied here.

After the implantation and the embedment take place successfully and the placental system starts its physiological activities in the maintenance of the embryo, which therefore becomes a self-sustained unit, more or less, it is expected that some kind of regularity should prevail in the mortality percentages. This is exhibited by the data from the 4th to the 8th month. Besides nutritional and excretional functions, the human placenta manufactures the hormones (gonadotropic hormone, oestrogen and progesterone) that are essential and sufficient for successful maintenance of the pregnancy to its full term, and the human embryo has not, therefore, to depend upon the pituitary, the ovary and the corpus luteum for these vital chemicals. (Arey, 1948; Windle, 1940; Hamilton, Boyd and Mossman, 1947; and Needham, 1931).

At and near the full term, again, when parturition must take place, certain hazards are involved in the process of birth. For instance, serious anatomical anomalies, if any, between the foetal size and the pelvic dimensions and the other difficulties of labour cause a number of death, which are absent during the earlier part of the pregnancy. The mortality at this stage, therefore, is expected to deviate from the regularity that is noticed between 4-8 months. Thus no regular mathematical law can be established to represent the mortality tendencies during the entire period of confinement.

We now proceed on with the practical method of mortality correction particularly of the observed U.S. data.

Strandskov and Siemens (1946) gave the numbers of males and females live-born, in single, twin and triplet confinements which are included in our table I. Now let n represent the number of confinements, male or female, single or multiple, n_l the number live-born out of n, and d the fraction of n lost by prenatal death. We have then the following relation :

$$n_l = n - n\,d\,;\ \text{or}\ n = \frac{n_l}{1-d}$$

Taking n_l from Table I and the corresponding d from Table III, corrected values of n i.e. m_1, f_1, m_2, f_2, etc. can be estimated.

The corrected values obtained in this way are given in Table IV, from which again, the corrected values of N_1, N_2, N_3, N, q_1, q_2, etc. have also been determined and included in the Table IV.

The corrected data, so obtained, are next used in the re-determination of μ, Φ, D, a, and other quantities. This has been carried out in the next section.

4. *Correction of the values of* μ, Φ, D, a, *and other quantities*:

Instead of re-calculating the values of the quantities, Φ, μ, D, etc., we shall adopt the method of infinitesimal calculus in estimating the amount of correction to be added (or subtracted) to the previous values.

We have seen above that m_1, f_1, m_2, f_2, etc., are all slightly increased after correction, above their respective values supplied by Strandskov and Siemens (1946). This is expressed in the differential notation by saying that their differentials are positive. A differential is negative when the variable decreases.

We now calculate the increments in the various quantities i.e. amount of correction to be added to them. We shall use the notation Δm_1, Δf_1, ΔN_2, etc., to represent the differentials of m_1, f_1, N_2 etc., respectively.

Thus, $\Delta m_1 = m_1$ (corrected) — m_1 (uncorrected) and similarly for the other differentials. The error, or the amount of correction expressed as p.c. of the uncorrected value is given by

$$(a) \quad \frac{\Delta m_1 \times 100}{m_1 \text{ (uncorrected)}}, \text{ and similarly for the others.}$$

And which is more important, the error, or the amount of correction expressed as a fraction of the uncorrected values is given by $\Delta m_1/m_1$ and so on for the rest.

We shall further assume that

$$(b) \quad \frac{\Delta r}{r} = \frac{\Delta m_2}{m_2} \text{ and } \frac{\Delta t}{t} = \frac{\Delta f_2}{f_2},$$

which must be true to the first order of approximation. Since, r and t represent respectively the frequencies of the 2-male and the 2-female twins, whereas m_2 and f_2 represent the number of

males and females respectively among the total twin confinements. We shall also assume that

$$(c) \quad \frac{\Delta r}{r} = \frac{\Delta r_1}{r_1} = \frac{\Delta r_2}{r_2} = \frac{\Delta m_2}{m_2} \text{ and } \frac{\Delta t}{t} = \frac{\Delta t_1}{t_1} = \frac{\Delta t_2}{t_2} = \frac{\Delta f_2}{f_2}$$

Since, r_1 and r_2 represent the frequencies of the 2-male twins, — the former monozygotic and the latter dizygotic, which have naturally the same prenatal mortality. Similarly, t_1 and t_2 represent the frequencies of the 2-female twins, — the former monozygotic and the latter dizygotic.

We now give below the formulae necessary to evaluate the errors in the various quantities :

(*d*) We have, $N_2 = r + s + t$

Hence, $\Delta s = \Delta N_2 - \Delta r - \Delta t$

i.e. $$\Delta s = \Delta N_2 - \frac{\Delta m_2}{m_2} r - \frac{\Delta f_2}{f_2} t,$$

by virtue of the relation (*b*) above.

(*e*) We have, $K_1 = \frac{1}{2} \frac{s r_1}{f_1 r_2}$, from Paper II, Section 14 (*b*)

Taking logarithmic differentials, we obtain

$$\frac{\Delta K_1}{K_1} = \frac{\Delta s}{s} + \frac{\Delta r_1}{r_1} - \frac{\Delta r_2}{r_2} - \frac{\Delta f_1}{f_1} = \frac{\Delta s}{s} - \frac{\Delta f_1}{f_1}$$

from the relation (*c*) of this section.

(*f*) Similarly, $K_2 = \frac{1}{2} \frac{s t_1}{m_1 t_2}$

Hence, $\frac{\Delta K_2}{K_2} = \frac{\Delta s}{s} - \frac{\Delta m_1}{m_1}$, also by virtue of (*c*) above.

(*g*) $K_1 = \mu (1 - \mu)$ from Part II, section 14 (*c*).

Hence $$\frac{\Delta K_1}{K_1} = \frac{\Delta \mu}{\mu} - \frac{\Delta \mu}{1 - \mu} = \frac{\Delta \mu}{\mu} \frac{(1 - 2\mu)}{1 - \mu}$$

Therefore, $$\frac{\Delta \mu}{\mu} = \frac{\Delta K_1}{K_1} (1 + \mu).$$

(*h*) $$\frac{\Delta\Phi}{\Phi}=\frac{\Delta K_2}{K_2}(1+\mu),\ \text{exactly as } (g).$$

(*i*) $$\text{Da}=\frac{\mu r_2}{r_1},\ \text{Hence, } \frac{\Delta D}{D}=\frac{\Delta\mu}{\mu}-\frac{\Delta a}{a}.$$

The value of Δa is obtained after determining the corrected value of '*a*' from the formula (*e*) in section 14, paper II.

Table IV embodies the corrected values of *r*, *s* and *t*, side by side, with those obtained from the raw data (Tables III & IV, Paper II). The corrected values of μ, Φ, *D*, *a*, and

TABLE IV

The raw data of Strandskow and Siemens, 1946, *and corrected for prenatal mortality, - 'whites'.*

QUANTITY	RAW DATA	CORRECTED DATA	QUANTITY	RAW DATA	CORRECTED DATA
m_1	14,249,501	14,340,311	r	108,772	110,473
f_1.	13,355,589	13,408,383	s	104,954	103,673
N_1	27,605,090	27,748,694	t	103,713	102,785
m_2	321,217	326,244	N_{22}/N_2	0.658440	0.658474
f_2.	309,243	312,036	$(m/f)_1$	1.066932	1.069503
N_2	315,230	319,140	$(m/f)_2$	1.038720	1.045533
m_3	4,565	4,656	$(m/f)_3$	1.001535	1.014379
f_3.	4,558	4,590	q_1	0.988600	0.988521
N_3	3,041	3,082	q_2	0.011289	0.011369
N.	27,923,410	28,070,916	q_3	0.000109	0.000110
t_1.	54,166	556,54	r_1	53,654	54,023

TABLE V

The theoretically calculated values of μ, Φ, D, a, and certain other quantities.

QUANTITY	FROM RAW DATA	FROM CORRECTED DATA	QUANTITY	FROM RAW DATA	FROM CORRECTED DATA
μ	0.003767	0.003798	$(m/f)_2$	1.039228	1.042617
Φ.	0.004070	0.004094	$(m/f)_3$	1.011835	1.010867
a	0.516115	0.516693	q_1	0.988562	0.988481
D.	0.007554	0.007608	q_2	0.011292	0.011370
a/b	1.066605	1.069078	q_3	0.000144	0.000146

TABLE VI.

Theoretical and observed values from the raw and the corrected U.S. Data on 'white' population.

QUANTITY	RAW DATA (U.S.)		CORRECTED DATA (U.S.)	
	Theoretical	Observed.	Theoretical	Observed.
$\left(\frac{m}{f}\right)_1$	1.039228	1.038720	1.042617	1.045533
$\left(\frac{m}{f}\right)_3$	1.011835	1.001535	1.010867	1.014379
q_1.	0.988562	0.988600	0.988481	0.988521
q_2.	0.011292	0.011289	0.011370	0.011369
q_3.	0.000144	0.000109	0.000146	0.000110
$m_1 : f_1$.	—	106.7 : 100	—	106.9 : 100
$m_2 : f_2$	103.9 : 100	103.9 : 100	104.3 : 100	104.6 : 100
$m_3 : f_3$	101.2 : 100	100.2 : 100	101.1 : 100	101.4 : 100

several other quantities calculated with the formulas of this section from the corrected data are given in Table V. It is found that the values of μ, Φ, D and a are not affected so much due to the corrections applied. They are increased only by 0.82%, 0.58%, 0.71% and 0.11% respectively. The triplet sex-ratio, of all quantities, is the most favourably affected and that is what was expected.

The closeness of agreement between the theoretical and the observed values of $(m/f)_2$, $(m/f)_3$, etc. is apparent from the results in Table VI. It shows that there is a satisfactory improvement in respect of $(m/f)_3$ which previously exhibited an appreciable difference between the theory and the observation.

In the last three rows of Table VI the sex-ratios have been put in thir popular forms. The other quantity, q_3 which exhibited an appreciable difference previously, does not yet show any betterment in the agreement. The refinement of the theory proposed to be done in the following sections will bring about certain changes in the various mathematical formulae deduced in Paper II in connection with the twin and the triplet statistics, and naturally, the expression for q_3 will undergo some modifications which are expected to improve the agreement between the calculated and the observed values of q_3.

It should, therefore, be emphasised here that the mortality error is not, in fact, solely or even primarily responsible for the discrepancy between the theory and the observation. The much-talked-of mortality factor has been given here the most careful and the best consideration in the interpretation of the data. The present author is of opinion that *it is the refinement of the theory, rather than the improvement of the data on the ground of prenatal mortality, that is more fundamentally important.*

B. Refinement of the Theory for Decline of the Twinning Processes with Increase of Plurality

5. *Introduction and biological basis.*

Prominent investigators on the problems of human twins seem to have realised the fact that the scission of a zygote (Dahlberg, 1926), which results in the formation of a monozygotic

twin pair can take place at any stage of development in the early life of the zygote or embryo, provided that the development or "organisation" does not exceed a certain ceiling level, beyond which the scission is impossible. It is further realised now-a-days that monovular twins with independent chorio-placentae are generated by the earliest scission of a zygote, and that the conjoined (Siamese) twins are the results of a belated twinning scission of a zygote (Price, 1950).

Norma Ford Walker (1952) has recently taken up this issue in connection with his study of the association of intra-pair differences in monozygotic twins with their twinning time. The following time-scale for twinning is quoted from his paper :

1. "Twinning of the embryonic cell mass occurring early before chorionic tissue has been established ; each forming its own chorion (2 placentae or 1 dichorionic placenta)."

2. "Twinning occurring after the embryonic cell mass is surrounded by the chorion ; a single chorion results with separate foetal circulations (monochorionic, separate foetal circulations)."

3. "Twinning occurring after both the chorion and the foetal circulation are developed ; a single chorion and a common circulation resulting (monochorionic, common foetal circulation)."

4. "Twinning of the embryonic cell mass delayed until both chorion and amnion have been formed ; the twins (either separate, or conjoined) being within a common amniotic sac (monochorionic, monoamniotic and a common foetal circulation)."

Obviously no scission is possible after the development of the embryo progresses beyond the stage (4). At present, we possess no precise knowledge about the durations of the above four stages in the development of the human embryo. Jenkins (1929) and recently, Dahlberg (1952) have envisaged the pos-

sibility of too early a scission of an ovum even before the reduction division, giving rise to two separate eggs and hence the formation of binovular twins. Whereas Jenkins introduced the idea in explaining the sex-homogeneity-index of triplets, Dahlberg has put forward this hypothesis to interpret some correlation observed between monovular and binovular twinning. So far as the monozygotic twins are concerned, such premature scissions of ova before the reduction division have no importance. We, therefore, take the scission of fertilised ova alone into our consideration here.

One very important point should be made clear in this connection, that monozygotic twinning scission may not always mean a complete division of the embryonic formative material into two independent masses which develop separately as two complete embryos. Separate establishment of two "organising centres" without any overlapping between their influences in a single formative area also leads to the formation of monozygotic twins. Whenever the scission of the formative embryonic material, or the separation of the spheres of influence of the "organising centres" are incomplete, "united twin monsters of the Siamese variety result" (Hamilton, Boyd and Mossman, 1947). Studies of monostrosities of various kinds indicate that the monozygotic twinning separation can take place along one of a number of possible axes (Gedda, 1951).

The four classes of monozygotic twins depending upon the time of actual manifestation of the inherent twinning tendency of a zygote clearly prove that the inherent twinning tendency in a zygote alone is not sufficient for causing the scission of the zygote. Certain suitable environmental conditions surrounding the zygote are also essential, which perhaps develop as a matter of physiological variation and are thus not always present. It may be, that all mothers do sometimes develop such suitable conditions, but in most of the cases, due to absence of zygotes having the twinning tendency, no monozygotic twins are formed. The possibility is also there, that mothers having the hereditary twin producing disposition alone, can sometimes develop the suitable conditions physiologically, in which the particular kind of zygotes, inherently possessing the twinning tendency, success-

fully form monovular twins. Whatever may be the case, it is, however, definite that the zygotes having the twinning tendency have to wait for an opportune moment for scission. If the opportunity comes very late, conjoined twins result, and if it is too late, no twins are formed. It is on this idea that the following mathematical formulations are based. The cause for early or late scission does not affect these formulations.

6. *Probability of the second monozygotic scission.*

Let p be the probability of the first monozygotic scission of a zygote or embryo and p', the probability of a similar scission of one of the two zygotes or embryos formed as a result of the first twinning scission of the initial zygote. We shall henceforth term the initial single zygote as the primary and each of the twins formed by its scission as the secondary zygote or embryo.

In consonance with what have been discussed in the preceding section, we suppose that there is a time-limit after the fertilisation of the ovum, within which the twinning process should commence, otherwise no twinning, completely separate or conjoined, can occur.

Let T be the interval of time between the formation of the zygote and the said time-limit. The monozygotic scission of the zygote can take place at any instant within the interval T. Let t be the time that actually elapsed before the scission of the primary zygote after its formation, so that the pair of secondary zygotes so formed are already somewhat old and advanced compared with the condition of the primary zygote at its initial state. The secondary zygotes cannot therefore have the same opportunity for twinning as the primary zygote had. The secondary zygotes have the time interval $(T\text{-}t)$ at their disposal within which only they can undergo twinning. This means that the probability of monozygotic scission of the secondary zygotes is proportionately less than that of the primary zygote, i.e., the probability should be $p\frac{T-t}{T}$.

Now t may vary from O to T, hence the average value of the probability (p') of monozygotic scission of the secondary zygotes is given by

$$p' = \frac{\int_0^T p \cdot \frac{T-t}{T} \cdot dt}{\int_0^T dt}$$

$$= \frac{1}{T^2} \int_0^T (T-t)\, dt$$

$$= \frac{1}{2} p .$$

Therefore, we find that *the probability of the second twinning scission is* $\frac{1}{2} p$, *and not* p *as hiterto supposed.*

7. *Probability of the third and the subsequent twinning scissions.*

As before, we suppose that the first scission occurs at time t after the formation of the primary zygote, and let us further suppose that the second scission takes place at time t' after the occurrence of the first scission. The average value of the probability of the third scission is, therefore, given by

$$p'' = \frac{p}{T} \frac{\int_0^T \int_0^{T-t} (T-t-t')\, dt' \cdot dt}{\int_0^T \int_0^{T-t} dt' \cdot dt} ,$$

since t' may vary from O to $(T-t)$ and t may vary from O to T. On integration, we get

$$p'' = \frac{1}{3} p .$$

Extending the argument to subesquent scissions, it can be proved that for the fourth scission, the probability is $\frac{1}{4}p$. Generalising from these results, we may say that the probability of the n/th scission is $(1/n)$ times the probability of the first scission.

8. *Modification of the twin formulae.*

Reference should be made to sections 13 onwards in paper (I) of the present series for the original forms of the various expressions and the basis of their derivations. We shall give below the modified forms of those expressions with brief explanations.

(a) *Monozygotic twins* :

The No. of 1-egg 2-male pairs $= Na\mu(1-\mu/2)^2\ (1-D)$ and the No. of 1-egg 2-female pairs $= N\,b\,\Phi\,(1-\Phi/2)^2\ (1-D)$, since, after the first scission is over, the probabilities of the second scission i.e., the scission of a secondary zygote or embryo are $\mu/2$ and $\Phi/2$ respectively for the male and the female sexes. Hence, the probability that for neither of the two secondary zygotes or embryos undergo such scissions are $(1-\mu/2)^2$ and $(1-\Phi/2)^2$ respectively for the male and the female sexes. The expressions in (*a*) admit of some simplifications on the approximations :

$$(1-\mu/2)^2 = 1-\mu \quad \text{and} \quad (1-\Phi/2)^2 = 1-\Phi\,,$$

As μ and Φ are small quantities, these approximations can be allowed here. Thus, we have,

(*b*) No. of 1-egg 2-male pairs $= N\,a\,\mu\,(1-\mu)\,(1-D)$ and No. of 1-egg 2-female pairs $= N\,b\,\Phi\,(1-\Phi)\,(1-D)$.

Hence,

$$N_{12} = N\,(1-D)\,[a\,\mu\,(1-\mu) + b\,\Phi\,(1-\Phi)]$$

(*c*) The sex-ratio among 1 egg twins is given by,

$$\left(\frac{m}{f}\right)_{12} = \frac{\mu}{\Phi} \cdot \frac{a\,(1-\mu)}{b\,(1-\Phi)} = \frac{\mu}{\Phi} \cdot \left(\frac{m}{f}\right)_{1}$$

As $$\Phi > \mu\,, \quad \left(\frac{m}{f}\right)_{12} < \left(\frac{m}{f}\right)_1$$

(*d*) *The dizygotic twins*:

$r_2 =$ No. of 2-egg 2-male pairs $= N\,a^2\,D\,(1-\mu)^2\,(1-D')$

$s =$ No. of 2-egg 1-male + 1-female pairs

Therefore,

$$s = 2\,Nab\,D\,(1-\mu)\,(1-\Phi)\,(1-D')$$

$t_2 =$ No. of 2-egg 2-female pairs $= N\,b^2\,D\,(1-\Phi)^2\,(1-D')$. Here, D' is taken as the probability that an additional zygote is formed after the formation of the first additional zygote. Hence, $(1-D')$ is the probability that no third independent zygote is at all formed. The exact relationship between D and D' cannot be given nor guessed at the present moment. The issue will be taken up again in a later section. The sex-ratio among the dizygotic twins remains unaffected. We have now,

$$N_{22} = N\,D\,(1-D')\,(1-a\,\mu-b\,\Phi)^2, \text{ since, } N_{22} = r_2 + s + t_2\,.$$

(*e*) The sex-ratio of all twin confinements is given by,

$$\left(\frac{m}{f}\right)_2 = \left(\frac{m}{f}\right)_1 \cdot \frac{\mu\,(1-D) + D\,(1-D')\,(1-a\,\mu-b\,\Phi)}{\Phi\,(1-D) + D\,(1-D')\,(1-a\,\mu-b\,\Phi)} < \left(\frac{m}{f}\right)_1,$$

(*f*) The relative frequencies of the 1-egg and the 2-egg twins remain unaffected by the introduction of the present modifications. Hence, the same formula as was given in section 19, part I of this series of papers for N_{22} is still valid. The form of the Weinberg's differential rule also remains just as before.

(*g*) Total number of twin pairs, N_2 is given by

$$\frac{N_2}{N} = q_2 = [a\mu(1-\mu) + b\Phi\,(1-\Phi)]\,(1-D) + D(1-D')\,(1-a\mu-b\Phi)^2.$$

9. *Modified forms of a few other formulae, and formulae for the evaluation of* μ, Φ, D, D', a, *etc.*

(i) From the analytical relations for m_1, f_1, r_1 and t_1 we have now very simple formulae for μ and Φ. Thus,

$$\mu = \frac{r_1}{m_1} \quad \text{and} \quad \Phi = \frac{t_1}{f_1}$$

Indications were given as to the procedure for estimating the values of r_1 and t_1 in section 21, paper I. It was given there that

$$r_1 = r - \frac{1}{2}\left(\frac{m}{f}\right)_1 s \quad \text{and} \quad t_1 = t - \frac{1}{2}(f/m)_1 s, \text{ where } r, s$$

and t are known from the twin confinement statistics.

(ii) $$\left(\frac{m}{f}\right)_1 = \frac{a(1-\mu)}{b(1-\Phi)}\Bigg), \text{ as before.}$$

This relation is most convenient for the determination of a and b, remembering that $a + b = 1$.

(iii) We have, instead of what were given in Section 21 (xiv), paper I, the relations,

$$\frac{r_1}{r_2} = \frac{(1-D)\cdot\mu}{a\,D\,(1-\mu)\,(1-D')} \quad \text{and} \quad \frac{t_1}{t_2} = \frac{(1-D)\cdot\Phi}{b\,D\,(1-\Phi)\,(1-D')}.$$

(iv) Therefore,

$$D\cdot\frac{1-D'}{1-D} = \frac{\mu}{1-\mu}\cdot\frac{r_2}{r_1} + \frac{\Phi}{1-\Phi}\cdot\frac{t_2}{t_1}.$$

Knowing μ and Φ and taking r_1, r_2, etc., from the twin statistics, D can be evaluated to the first approximation on the assumption that $D' = D$ in $1 - D'$.

In justification of the above procedure, a few words are added here : As μ, Φ, etc., on the right-hand-side of the above formula are directly obtained from the statistical data, there is no error in them through approximations. Hence, it is easily seen by logarithmic differentiation that,

$$\frac{d\,D}{D} + \frac{d\,D}{1 - D} = \frac{d\,D'}{1 - D'} ,$$

in which D and D' are negligibly small compared to 1.

Hence,

$$\frac{d\,D}{D} = D' \left(\frac{d\,D}{D'} \right) .$$

Therefore, percentage error in D is D' times the percentage error in D'. As D' must be a very small fraction, it follows that no large error is introduced in D, due to an error in D' by approximation. The above relation, therefore, can be used as a formula for evaluating D, assigning to D' any reasonable approximate value.

(v) An alternative method, free from such approximation and much easier than the one explained in (iv) is to evaluate D through the relation,

$$\frac{N_1}{N} = q_1 = (1 - D)\,(1 - a\,\mu - b\,\Phi) . \quad \text{(Section 14, paper I).}$$

Since, q_1, a, b, μ, and Φ are not subject to any error due to approximation, D can be reliably computed out through this relation.

(vi) The theoretical values of $(m/f)_2$ and q_2 can be determined though the formulas 8 (e) and 8 (g) of the preceding section.

(vii) The expressions for the quantities r, s, and t which were supposed to represent the frequencies of the sex-combinations, 2-male, 1-male + 1-female, and

2-female among the twin confinements are transformed into the following forms after the refinement) (See sec. 19, paper I and sec. 14, paper II in this connection).

$$r = N\,a\,(1-\mu)\,[\mu\,(1-D) + a\,D\,(1-\mu)\,(1-D')]\,,$$

$$s = 2\,N\,a\,b\,D\,(1-\mu)\,(1-\Phi)\,(1-D'),\ \text{and}$$

$$t = N\,b\,(1-\Phi)\,[\Phi\,(1-D) + b\,D\,(1-\Phi)\,(1-D)].$$

Therefore,

$$r : s : t = a\,(\mu + a\,D) : 2\,a\,b\,D : b\,(\Phi + b\,D)\,,$$

on neglecting μ, Φ, D, D' in comparison with unity. Assuming further, $a = b$, and $2\mu = 2\Phi = D$, we get, $r : s : t = 1 : 1 : 1$ as before (sec. 18, paper I).

The frequencies r, s, t, or their ratios can be estimated through the formulae given above and compared with those obtained from the statistical data.

10. *Modifications of the Triplet formulae.*

We have supposed, in course of our analyses, in the preceding sections that the probability of the second zygote being added to the first or the primary zygote is D, whereas that for the third being added to what is already a dizygotic twin pair is D' and that D' is less than D. But how much less is a question, the answer to which is deferred to a latter section. Let us assume in the same way, that D'' is the probability for the fourth zygote being added to what is already a trizygotic triplet set and that D'' is less than D', and so on.

In the case of twinning scission of a zygote or embryo, we have proved already that the probabilities of the first, the second, the third and the subsequent scissions for a male primary zygote with inherent twinning tendency are μ, $\mu/2$, $\mu/3$, and so on respectively. For a female zygote the same probabilities are Φ, $\Phi/2$, $\Phi/3$, and so on. The introduction of these modified values for the probabilities of subsequent twinning, ne-

cessarily, causes certain important changes in the forms of the triplets formulae, although the genesis of the triplets remains the same as before.

(*a*) 1-*egg Triplet sets.* (*see sec.* 6, *paper II*) :

$$\text{No. of 1-egg 3-male sets} = 2\,N\,a/\mu\,2 \cdot (1-\mu/2)\,(1-\mu/3)^2\,(1-D)$$
$$= N\,a\,\mu^2\,(1-\mu/2)\,(1-\mu/3)^2\,(1-D)\,.$$

Here $(1-\mu/2)$ stops the second scission and hence also subsequent scissions of one of the member of the twin pair formed by the first twinning division and $(1-\mu/3)^2$ stops the subsequent twinning scissions, after the second, of the pair formed by the scission of the other member of the twin pair generated by the first twinning division of the primary zygote. No detailed explanations will be given for the derivations of the various expressions and the probability functions used therein for the sake of brevity.

Similarly,

$$\text{No. of 1-egg 3-female sets} = N\,b\,\Phi^2\,(1-\Phi/2)\,(1-\Phi/3)^2\,(1-D).$$

Neglecting μ^2 and Φ^2, and their still higher powers compared to 1, we obtain,

$$\text{No. of 1-egg 3-male sets} = N\,a\,\mu^2\,(1-7\,\mu/6)\,(1-D)\,,$$
$$\text{No. of 1-egg 3-female sets} = N\,b\,\Phi^2\,(1-7\,\Phi/6)\,(1-D)\,.$$

From a knowledge of N, a, μ, Φ and D, these frequencies can easily be determined through the above formulas.

(*b*) Again, $$q_{13} = \frac{N_{13}}{N} = (1-D)\left[(a\,\mu^2 + b\,\Phi^2) - \frac{7}{6}\,(a\,\mu^3 + b\,\Phi^3)\right]$$

It is important to note that N_{13} and q_{13} are now about half their previous values given by the expressions in section 6, paper II.

$$\left(\frac{m}{f}\right)_{13} = \frac{a\,\mu^2\,(1-7\,\mu/6)}{b\,\Phi^2\,(1-7\,\Phi/6)} \cdot \frac{a\,\mu^2\,(1-\mu)\,(1-\mu/6)}{b\,\Phi^2\,(1-\Phi)\,(1-\Phi/6)}$$

Hence,

$$\left(\frac{m}{f}\right)_{13} = \left(\frac{m}{f}\right)_{11} \frac{\mu^2 (1 - \mu/6)}{\Phi^2 (1 - \Phi/6)} = \left(\frac{m}{f}\right)_{12} \frac{\mu (1 - \mu/6)}{\Phi (1 - \Phi/6)}$$

Therefore,

$$\left(\frac{m}{f}\right)_{13} < \left(\frac{m}{f}\right)_{12} < \left(\frac{m}{f}\right)_{11}$$

Again,

$$\left(\frac{m}{f}\right)_{13} : \left(\frac{m}{f}\right)_{12} = \frac{\mu}{\Phi} \cdot \frac{(1 - \mu/6)}{(1 - \Phi/6)} \fallingdotseq \frac{\mu}{\Phi},$$

and

$$\left(\frac{m}{f}\right)_{12} : \left(\frac{m}{f}\right)_{11} = \frac{\mu}{\Phi}.$$

These relations, therefore, agree with what were proved in sec. 6, paper II and are extremely interesting. The values of q_{13} as well as $\left(\frac{m}{f}\right)_{13}$ can be computed with the formulas given above, when μ, Φ, D and a are all known.

(*c*). *The 2-egg triplets*: (*see sec. 7, paper II*):

No. of 2-egg 3-male sets $= 2 N a^2 D \mu (1 - \mu)^2 (1 - D')$,

No. of 2-egg (2-male + 1-female) sets $= 2 N a b D \mu (1 - \mu) \cdot (1 - \Phi) (1 - D')$

No. of 2-egg (1-male + 2-female) sets $= 2 N a b D \Phi (1 - \mu) \cdot (1 -) (1 - D')$

No. of 2-egg 3-female sets $= 2 N b^2 D \Phi (1 - \Phi)^2 (1 - D')$.

We have assumed that $(1 - \mu/2)^2 = (1 - \mu)$

and that $(1 - \Phi/2)^2 = (1 - \Phi)$.

Hence,

$$q_{23} = N_{23}/N = 2 D (1 - D') (1 - a\mu - b\Phi) [a\mu (1 - \mu) + b\Phi (1 - \Phi)].$$
$$= 2 D (1 - D') (1 - a\mu - b\Phi) [(a\mu + b\Phi) - (a\mu^2 + b\Phi^2)].$$

Each of the above frequencies including N_{23} and the quantity q_{23} can be evaluated through these formulae from a knowledge of μ, Φ, D, a and N.
Again,

$$\left(\frac{m}{f}\right)_{23} = \left(\frac{m}{f}\right)_{11} \cdot \frac{3\,a\,\mu\,(1-\mu) + b\,(1-\Phi)\,(2\,\mu+\Phi)}{3\,b\,\Phi\,(1-\Phi) + a\,(1-\mu)\,(2\,\Phi+\mu)} < \left(\frac{m}{f}\right)_{11}$$

The sex-ratios among the monovular twins, the monovular as well as the binovular triplets are, therefore, each less than the single confinement sex-ratio, i.e., the former are less biased in favour of males than the latter. It is proved below, however, that the sex-ratios among the single confinements, the binovular twins and the tri-ovular triplets are equal to one another.

(*d*). *The 3-egg triplets*: (*see sec.* 8, *paper II*):

No. of 3-egg 3-male sets $= N\,a \cdot D\,a \cdot D'\,a\,(1-D'')\,(1-\mu)^3$
$= N\,a^3\,D\,D'\,(1-D'')\,(1-\mu)^3.$

No. of 3-egg (2-male + 1-female) sets $= 3\,N\,a^2\,b\,D\,D'\,(1-D'') \cdot$
$\cdot\,(1-\mu)^2\,(1-\Phi)$

No. of 3-egg (1-male + 2-female) sets $= 3\,N\,a\,b^2\,D\,D'\,(1-D'') \cdot$
$\cdot\,(1-\mu)\,(1-\Phi)^2$

No. of 3-egg 3-female sets $= Nb^3DD'\,(1-D'')\,(1-\Phi)^3$

Again,

$$q_{33} = N_{33}/N = (1 - a\,\mu - b\,\Phi)^3\,DD'\,(1 - D'')$$

And,

$$\left(\frac{m}{f}\right)_{33} = \left(\frac{m}{f}\right)_{11},$$

from the above frequencies.

Therefore,

$$\left(\frac{m}{f}\right)_{11} = \left(\frac{m}{f}\right)_{22} = \left(\frac{m}{f}\right)_{33}$$

The frequencies of the sex-combinations of 3-egg triplets and the value of q_{33} can be calculated through the help of the above formulae, when μ, Φ, a, D, D' and D'' are all known.

(*e*) The total triplet frequency N_3 is given by the relation,

$$\frac{N_3}{N} = q_3 = q_{13} + q_{23} + q_{33} .$$

Hence from (*b*), (*c*) and (*d*) of this section,
We obtain,

$$q_3 = (1 - D)\left[(a\mu^2 + b\Phi^2) - \frac{7}{6}(a\mu^3 + b\Phi^3)\right.$$
$$+ 2D(1 - D')(1 - a\mu - b\Phi)\left[(a\mu + b\Phi) - (a\mu^2 + b^2)\right]$$
$$+ DD'(1 - D'')(1 - a\mu - b\Phi)^3 .$$

Therefore, q_3 can be precisely evaluated when D, D', D'', μ, Φ, and a are all known. (compare this refined expression for q_3 with the previous one in section 14 (i), paper II).

(*f*) The triplet sex-ratio is $(m/f)_1$ times a fraction, of which the numerator is

$$3\mu^2(1 - \mu/6)(1 - D) + 2D(1 - D') \cdot$$
$$\cdot [3a\mu(1 - \mu) + 2b\mu(1 - \Phi) + b\Phi(1 - \Phi)] +$$
$$+ 3DD'(1 - D'')(1 - a\mu - b\Phi)^2 ,$$

and the denomenator is

$$3\Phi^2(1 - \Phi/6)(1 - D) + 2D(1 - D') \cdot$$
$$\cdot [3b\Phi(1 - \Phi) + 2a\Phi(1 - \mu) + a\mu(1 - \mu)] +$$
$$+ 3DD'(1 - D'')(1 - a\mu - b\Phi)^2 .$$

The sex-ratio for triplets can be calculated from the formula given here and compared with that actually observed. The above expressions for q_3 and $\left(\frac{m}{f}\right)_3$ can be simplified on neglect-

ing terms involving μ, Φ, D, etc. in higher than the third degree. But for the purpose of practical computations, the forms presented here will be found quite convenient.

11. *Relative frequencies of the monozygotic, the dizygotic and the trizygotic triplet sets*:

As in section 9, paper II, we suppose that U, V, W, and X represent the frequencies of the four sex-combinations, (i) 3-male, (ii) 2-male + 1-female, (iii) 1-male + 2-female, and (iv) 3-female respectively. The distribution of these sex-combinations over the three types of triplets arising out of one egg, two eggs and three eggs is shown in table VII, which differs most remarkably from the previous distribution so far as the 1-egg triplet sets are concerned. The relative frequencies given in table VII are obtained after certain simplifying approximations which, however, do not affect the results so appreciably. For example, it has been assumed that μ, Φ, D, D', etc., are negligibly small compared to unity. This assumption is supported by the actual numerical values obtained for the quantities.

Expressions for the frequencies N_{13}, N_{23} and N_{33} can be written down in terms of the frequencies U, V, W and X with the help of this chart, exactly in a manner followed in section 9, paper II. We shall return to this discussion in connection with the U.S. data in a latter section of this paper.

TABLE VII

Distribution of U, V, W and X over 1; 2 and 3-egg triplet sets.

TOTAL FREQUENCY OF SEX-COMBINATIONS	RELATIVE FREQUENCIES		
	1-egg	2-egg	3-egg
U.	μ^2	$2a\mu D$	$a^2 D D'$
V.	0	2μ	$3a D'$
W	0	2Φ	$3b D'$
X.	Φ^2	$2b\Phi D$	$b^2 D D'$

12. *The Sex-homogeneity Index*:

This subject was treated in section 12, paper II and reference was made to the works of Jenkins (1929) and Dahlberg (1926) in that connection.

The expressions for the sex-homogeneity index (S.H.I.) after the refinements discussed above stand as follows, on neglecting μ, Φ, D, etc. compared to unity:

$$S.H.I. = \frac{U+X}{V+W} = \frac{a(\mu+aD)^2+b(\Phi+bD)^2-D^2(1-D'/D)(1-3ab)}{2\,a\,b\,D\,(\mu+\Phi+3/2\,.\,D')}.$$

In the case of U.S. data, we have proved in a latter section that $D' = \frac{2}{3} D$ for the 'white' population. Hence,

$$S.H.I. = \frac{a\,(\mu+a\,D)^2+b\,(\Phi+b\,D)^2-D^2\left(\frac{1}{3}-a\,b\right)}{2\,a\,b\,D\,(\mu+\Phi+D)}.$$

Assuming further, $\Phi = \mu$, and $a = b = \frac{1}{2}$, we obtain

$$S.H.I. = \frac{\mu^2+\mu\,D+\frac{1}{6}D^2}{\mu\,D+\frac{1}{2}D^2}.$$

If again we put $D = 2\,\mu$ which is very nearly the case with U.S. data and roughly wherever the frequency of dizygotic twin pairs is about double that of the monozygotic pairs, we obtain

$$S.H.I. = \frac{11}{12} = 0.916,$$ a value agreeing very

satisfactorily with those actually obtained by the various authors. Jenkins (1929) quoted the observed values of the sex-homogeneity-indices for a few countries as follows: Prussia — 0.876, Italy — 0.968, Austria — 0.985, Hungary — 0.701, Germany — 0.923 and Belgium — 0.958, whereas the computed values from the

twin data of France and Denmark were given as 0.691 and 0.676 respectively. There was, therefore, a wide difference between the observed and the theoretical results. The value 0.916 given by the theory presented here, of course, agrees closely with the above observed values, except that for Hungary. It should be mentioned here that the relative proportions of the binovular and the monovular twin pairs in European countries and in U.S. are nearly equal and most probably that accounts for the nearly equal values for the S.H.I. in these countries.

13. (*a*) *Determination of D′ and D*:

In sub-sections 9 (iv) and 9 (v) we have referred to two methods for the determination of D. We have from 10 (*e*),

$$q_3 = q_{13} + q_{23} + q_{33}, \qquad \ldots (1)$$

Where,

$$q_{13} = (1 - D)\left[(a\mu^2 + b\Phi^2) - \frac{7}{6}(a\mu^3 + b\Phi^3)\right. \qquad \ldots (2)$$

$$q_{23} = 2D(1 - D')(1 - a\mu - b\Phi)\left[(a\mu + b\Phi) - (a\mu^2 + b\Phi^2)\right. \ldots (3)$$

$$q_{33} = D\,D'(1 - D'')(1 - a\mu - b\Phi)^3. \qquad \ldots (4)$$

We have also,

$$(1 - D)\,N_{22}/N_1 = D(1 - D')(1 - a\mu - b\Phi) \qquad \ldots (5)$$

from section 14 & 15, paper I.

Now eliminating the factor $(1 - D')$ between (3) and (5), we have

$$q_{23} = 2(1 - D)\frac{N_{22}}{N_1}\left[(a\mu + b\Phi) - (a\mu^2 + b\Phi^2)\right]\ldots \qquad \ldots (6)$$

Thus q_{13} and q_{23} are now both independent of D', which can, therefore, be evaluated from known values of a, b, μ, Φ, and D, obtained through the formula 9 (v). As q_3 is directly available from the confinement statistics, q_{33} can now be found out. Equation (4) can now be used for the determination of D', assuming D'' to be equal to D in $(1 - D'')$. By the logarithmic differential method used in subsection 9 (iv), it can

be shown that this approximation in the value of D'' shall not affect the result here to any large degree. On obtaining the value of D', we may, as a second approximation, put D' for D'' and re-calculate the value of D' and repeat the process twice or thrice and obtain a more accurate value of D'.

(*b*) The value of q_3 can, obviously, be estimated after D and D' are evaluated as indicated above. As D'' cannot be greater than D' and as a slight error in D'' does not affect the value of $(1-D'')$ to any appreciable extent, we can substitute the value of D' for D'' in $(1-D'')$ in (4).

With the same assumptions, the sex-ratio, $\left(\frac{m}{f}\right)_3$, can be ascertained through the formula 10 (*f*).

The practical method of computations are illustrated in the following sections.

14. *Calculation of*, μ, Φ, a, b, D, D', *etc.*

The actual procedure to follow in the mathematical analyses of the confinement statistics is demonstrated in this section with the U.S. data of the 'white' population, corrected for prenatal mortality.

(*a*) From section 9 (i), we have

$$\mu = \frac{r_1}{m_1} = \frac{54{,}302}{14{,}340{,}311} = \cdot 003786$$

$$\Phi = \frac{t_1}{f_1} = \frac{54{,}655}{13{,}408{,}383} = \cdot 004076 .$$

The values of r_1, t_1, m_1 and f_1 are taken from the columns 3 and 6 of table IV, which are corrected for prenatal mortality. Where such corrections are not available the raw data can be used exactly as above. Using the raw data from the columns 2 and 5 of the same table, we get $\mu = 0.003752$ and $\Phi = 0.004055$, which deviate from the above results by 0.9 % and 0.5 % respectively. Comparing the results of this section with those in Table V, it follows that μ and Φ are not affected

to any appreciable degree by the introduction of the two kinds of refinements. This does not, however, mean that μ and Φ do not require carefully collected data for their correct detremination. As μ and Φ are each a ratio of two quantities which are almost similarly affected by mortality and as $(1 - \mu)$ and $(1 - \Phi)$ are very nearly equal to unity, the two refinements do not affect the results materially. But whatever modifications are effected should be given due importance and the basic data must be derived from reliable statistics.

(*b*) The formula 9 (ii) is now applied to find out a and *b*. We have

$$\left(\frac{m}{f}\right)_1 = \frac{a\,(1-\mu)}{b\,(1-\Phi)} \text{ and } a+b=1\text{, where } \left(\frac{m}{f}\right)_1 = 1.069503$$

from table IV, and μ and Φ are known from (*a*) above. Thus $\frac{a}{b} = 1.069186$, and hence, $b = 0.483282$ and $a = 0.516718$. Comparing their previous values in table V, it is found that *a* and *b* also are not appreciably altered by the refinements.

(*c*) The formula 9 (v),

$$q_1 = (1 - D)\,(1 - a\,\mu - b\,\Phi)$$

is used in calculating *D*. We have, therefore,

$$1 - D = \frac{0.988521}{0.996075} = 0.992416$$

Hence, $D = 0.007584$, using q_1 from column 6 of table IV, and *a* and *b* as determined above. The values of *D* as given in table V, and deduced from the raw data and the data after the mortality correction are 0.007554 and 0.007608 respectively.

(*d*) Determination of D' depends upon the statistical value of q_3 which, in our example, is 0.000110, and the theoretically calculated values of q_{13} and q_{23}.

Thus,

$$q_{13} = (1 - D)\left[(a\mu^2 + b\Phi^2) - \frac{7}{6}(a\mu^3 + b\Phi^3)\right] \text{ from 10 } (e).$$

$$= 0.992416 \times \cdot 00001542 = \cdot 00001531.$$

$$q_{23} = 2(1 - D)\frac{N_{22}}{N_1}[(a\mu + b\Phi) - (a\mu^2 + b\Phi^2)] \text{ from 13 } (a) \text{ (6)}$$

$$= 2\,(.992416)\,(.007573)\,(.003910)$$

$$= 0.00005877.$$

since,

$$N_{22} = 210{,}145 \text{ and } N_1 = 27{,}748{,}694 \text{ from table IV.}$$

Hence,

$$q_{33} = q_3 - q_{13} - q_{23} = .00003592.$$

From 13 (*a*) (4),

$$q_{33} = D\,D'\,(1 - D'')\,(1 - a\mu - b\Phi)^3$$

$$= (.007584)\,D'\,(.992416)\,(.988271),$$

assuming as a first approximation, $D'' = D$ in $1 - D''$.

Therefore,

$$D' = \frac{.00003592}{.00743821} = 0.004829 \text{ (approx.)}$$

and

$$1 - D' = 0.995171.$$

Substituting this value of D' for D'' in $1 - D''$ and recalculating, we obtain, as a nearer approximation $D' = 0.004829$, i.e., *D' is about $\frac{2}{3}$ of D.*

It is no wonder that the latter estimate of D' does not deviate from the former, and it justifies the assumption made in substituting D for D'' in $1 - D''$.

(*e*) The formulae given in 8 (*e*) and 8 (*g*) are now applied for the computation of $\left(\frac{m}{f}\right)_2$ and q_2 respectively.

We have $\left(\frac{m}{f}\right)_1 = 1.069503$ according to corrected data.

i.e. About 107 males to 100 females in single confinement. Therefore,

$$\left(\frac{m}{f}\right)_2 = (1.069503)\ (.974742) = 1.042489\ ,$$

i.e. Twin sex-ratio is 104.2 males to 100 females.

And, $q_2 = 0.011365 = 1/87.98$.

i.e. The frequency of twin confinements $= \frac{1}{87.98} \times$ the frequency of all confinements.

This corresponds to the Hellin's Law. The observed twin sex-ratio after mortality correction is 104.6 males to 100 females, which agrees well with the theoretical value, 104.2.

Another interesting fact may be seen through the formula 8 (*e*) for $\left(\frac{m}{f}\right)_2$. Neglecting D, D', μ and Φ compared to unity, we obtain,

$$\left(\frac{m}{f}\right)_2 \Big/ \left(\frac{m}{f}\right)_1 = \frac{\mu + D}{\Phi + D} \cdot = 0.975128$$

The ratio of the observed twin and single confinement sex-ratio is 0.977584, which value departs but by about 0.33 % from the ratio, $(\mu + D)/(\Phi + D)$, calculated above.

Hence, the above simplified relation between the two sex-ratios may be used with fair accuracy instead of the formula 8 (*e*) which is rather complex, for evaluating $\left(\frac{m}{f}\right)_2$

(*f*) From the expression for $\left(\frac{m}{f}\right)_3$ given in 10 (*f*), the sex-ratio for the triplet confinement can be calculated.
Thus,

$$\left(\frac{m}{f}\right)_3 = \left(\frac{m}{f}\right)_1 \times (0.952008) = (1.069503)\ (0.952008) = 1.020680$$

i.e. the theoretical triplet sex-ratio is 102.1 males to 100 females.

The corrected observed ratio is 101.4 males to 100 females. The agreement between the two results is therefore satisfactory.

(*g*) From the manner of determination of D' in 14 (*d*), it is obvious that the calculated value of q_3 will agree perfectly with the observed value. The statistical value of q_3 has been used in the determination D'. This means that D' has been so selected that it fits exactly with the observed data.

There is, in the present case, another alternative method of evaluating D' quite independent of the triplet data. Thus, we have in 13 (*a*) (5),

$$(1 - D)\frac{N_{22}}{N_1} = D(1 - D')(1 - a\mu - b\Phi)$$

None of the quantities involved here depend upon the triplet data for their evaluation. Substituting the values of a, b, μ, Φ, D from this section and those of N_{22} and N_1 from the U.S data for 'white' population from table IV, we obtain $D' = 0.005118$ and $D'/D = 0.6749 \approx 2/3$. This estimate of D' is of the same order of value as the one found out in (*d*) above.

Even if this value of D' which is determined independent of triplets data is used in calculating the value of q_3, the theoretical value will agree with the observed value quite closely. The present author, however, recommends the application of the method outlined in (*d*) of this section for the determination of D', in preference to the formula 13 (*a*) (5). For, by the method of logarithmic differentials illustrated in section 9 (iv), it can be proved that, $\frac{dD'}{D'} = \frac{1}{D'} \cdot \left(\frac{dD}{D}\right)$ in the case of the latter formula.

The p.c. error in D' is, therefore, $\frac{1}{D'}$ times the p.c. error in the estimation of D. As $1/D'$ is a very large quantity (about 200), this estimation of D' is liable to a large error, unless dD/D is vanishingly small.

The agreement between the two estimates of D' is then an indirect indication of the fact that the value of D, in this case, is free from any appreciable error, and thus the simpler formula for the determination of D' finds a justification.

The agreement now achieved between the theoretical and the observed values of q_3 should not, therefore, be regarded as artificial or forced. The value of q_3 obtained by putting $D' = .005118$ is 0.000112, as against the observed value of 0.000110. The agreement is close enough.

15. *The raw U.S. data for the 'white' and the 'colored' populations and the refined theory* :

We must critically examine the agreement between the refined theory and the raw U.S. data as supplied by Strandskov and his associates (1945, 1946). This only can properly show the importance of the mortality correction performed in chapter (A) of this paper. Strandskov's data of the registration area included still-births, which were, however, defective only in respect of the early abortions. In the scientific quarters this flaw in the confinement statistics has been taken as the root-cause for the higher female frequencies among plural confinements.

The present author has, therefore, given the fairest consideration to this question in the sections 1 to 4 of this paper. We have seen there that the results improve particularly in respect of the triplets sex-ratio, but no improvement in respect of q_3 was noticed. And that could mean that the slight error in the data which were fairly reliable, was not solely responsible for the observed discrepancies.

In table VIII have been arranged the results of the application of the refined theory presented in this paper to the raw U.S. data as published by Strandskov and his associats (1945 & 1946) and to those data rectified for the prenatal mortality (table IV). Attention may now be drawn also to table VI which is a comparative study of the results of application of the theory presented in papers I and II to the raw U.S. data as well as the data corrected for prenatal mortality.

In table VIII, the value of D' in the parentheses has been determined by the simple method shewn in 14 (g), whereas the other value has been obtained by the method explained in 14 (d). Values of other quantities, q_3 and *S.H.I.*, calculated with the former value of D' are also given in parentheses.

TABLE VIII

Refined theory in the interpretation of the Raw and the corrected U. S. data for the 'white'. (Strandskov '45 & '46)

QUANTITY	RAW U.S. DATA		CORRECTED U.S. DATA	
	Theoretical	Observed	Theoretical	Observed
$\left(\frac{m}{f}\right)_1$ . . .	1.066932	1.066932	1.069503	1.069503
$\left(\frac{m}{f}\right)_2$ . . .	1.039794	1.039228	1.042489	1.045533
$\left(\frac{m}{f}\right)_3$ . . .	1.015727	1.001535	1.020680	1.014379
q_1	0.988600	0.988600	0.988521	0.988521
q_2	0.011291	0.011289	0.011365	0.011369
q_3	(0.000115) 0.000109	0.000109	(0.000112) 0.000110	0.000110
μ	.003752	—	.003786	—
Φ	.004056	—	.004076	—
a	0.516115	—	0.516718	—
D	0.007530	—	0.007584	—
D'	(0.005199) 0.004851	—	(0.005118) 0.004829	—
$S.H.I.$ . . .	(0.927) 0.952	—	0.951	—

In table IX the results of applying the refined theory to the U.S. data for the 'colored' population (Strandskov, '45 and '46) are included and compared with those obtained previously in paper II (tables III & IV).

TABLE IX

The refined and the original theories in the interpretation of the U. S. data for the 'colored'.

QUANTITY	ORIGINAL THEORY	OBSERVED DATA	REFINED THEORY
$(m/f)_1$	1.054890	1.054890	1.054890
$(m/f)_2$	1.007401	1.008959	1.007760
$(m/f)_3$	0.964615	0.900000	0.984564
q_1.	0.985632	0.985643	0.985643
q_2.	0.014125	0.014155	0.014166
q_3.	0.000220	0.000197	0.000197
μ	0.003833	—	0.003818
Φ.	0.004500	—	0.004480
a	0.513191	—	0.513193
D.	0.010253	—	0.010256
D'	—	—	0.009620

16. *Comments*:

It is gratifying to note that the refined theory presented here accounts satisfactorily for all the observed confinement data for the U.S. 'white' population. The prenatal mortality correction though effects a closer agreement between the theoretically predicted and the observed results, yet it appears to have much less importance than the modification of the original theory (papers I & II) into its refined form.

The theory both original and the refined, however, seem to have failed in interpreting the extremely low observed value, 0.90, of the 'colored' triplet sex-ratio. The expected values are 0.965 and .985 approximately according to the original and the refined forms of the theory respectively. This is most probably due to large error in the observed value of the triplet sex-ratio itself. The total number of 'colored' triplet sets on which the data are based is only 703 as compared with the 'white' triplet sets numbering 3,041 and hence, the present meagre 'colored' triplet data are liable to large sampling errors. This contention is further supported from a study of the general trend of the sex-ratios among the single, the twin and the triplet confinements of the 'white' and the 'colored' U.S. populations. It is found that whereas in the case of the 'white' single, twin and triplet confinements and the 'colored' single and twin confinements, the sex-ratios show more or less a uniform decrease in their values with each increase in the number of embryos per pregnancy, the difference between the 'colored' triplet sex-ratio and the 'colored' twin sex-ratio is rather unusually abrupt and great. Actually, a value near about the expected value of 0.97 or 0.98 would fit in with the general trend of the sex-ratios mentioned above very normally.

The same reason may also be held responsible for the difference in the relative values of D', i.e. D'/D, for the two U.S. populations. Of course it is possible that the two different populations possessing different genetic compositions may have different values for the ratio D'/D. The biological implications of this ratio is too deep for immediate elucidation. It has a direct bearing upon the genesis of dizygotic twins and heredity of the dizygotic twinning tendency in the two populations.

It should finally be remarked that the data analysed here are not grouped according to the age of the productive mothers. Such grouping of the data as emphasised in the preceding papers (I & II), is very essential for strict analysis.

The U.S. data for the 'white' have been used here to demonstrate the steps which one has to follow in analysing a given mass of confinement statistical material.

The values of μ, Φ, D *and a may be regarded as ethnic constants which have different values for different age-groups of mothers.*

The determination of parameters such as μ, Φ, D and a from statistical materials is always liable to appreciable, or considerable errors unless the material, even collected under ideal conditions, is sufficiently large. Hence, in order to have characteristic values of these quantities for any ethnic group, a large volume of reliable data is absolutely necessary. In our illustrations here, the data for the U.S. 'white' satisfy this condition pretty well and thus reduce the error in D and other estimates to insignificant values. This further elucidates the point discussed in section 14 (*g*) in connection with the determination of D' from the simpler formula given there. The same formula when applied to the 'colored' data would yield an unacceptably high value ; because in the latter case the volume of the data is small and thus D and other estimates are subjected to appreciable or large sampling errors.

In passing, it may be pointed out that another alternative method of evaluating out D' is to use the formula for q_2 in section 8 (*g*). There, substituting q_2 from the observed data and the values of μ, Φ, a and D in the right-hand-side, the value of D' is readily obtained, quite independent of the triplet data. For the U.S. 'white' this estimate of D' is 0.004651 which is of the same order as the previous values in 14 (*g*). This further strengthens the concept of D'. In the case of the 'colored' data, D' comes out in this way to be 0.010712 which is even greater than D and hence, absurdly high, for reasons explained in the preceding paragraphs on sampling error.

It will be interesting now to apply the methods of analyses presented here to properly collected data of various populations of the world.

In that connection it would be of much interest to verify also the empirical rules, deduced in section 2, regarding the prenatal mortalities in the single and the plural confinements.

The present author will gratefully receive data from all quarters in this connection if they are generously communicated to him.

The author takes this opportunity to express his thanks to Professor C. Gini, University of Rome, for his constructive criticisms. He is thankful to Dr. B.S. Guha, the past Director, Department of Anthropology, Government of India, for his constant encouragement. Thanks are also due to Dr. E.C. Büchi who has carefully perused the manuscript, and to the statistical section, for helping me in doing a part of the numerical calculations.

Department of Anthropology, Government of India Indian Museum, Calcutta-13).

REFERENCES

1. Arey, L. B. 1948, *Developmental Anatomy*, p. 150. W. B. Saunders Company, Phila, and Lond.
2. Auerbach, E. 1912, « Arch. f. Rassen U. Gesselschafts Biol. » 9 : 10.
3. Barnett, Anthony 1950, *Human Species*, Macgibbon & Kee Ltd., London.
4. Blandau, R. J. and W. C. Young 1939 « Am. J. Anat. » 64 : 303.
5. Blandau, R. J. and E. S. Jordan 1941 « Am. J. Anat. » 68 : 276.
6. Boldrini, M. 1936, *La proportion des sexes dans les Conceptions humaines*, « Rev. Inst. Internat. Stat. » 4 : 484.
7. Ciocco, A. 1940, *Sex differences in Morbidity and Mortality*, « Quarterly Rev. of Biol. » 15 : 59-73 & 192-210.
8. Dahlberg, Gunnar 1926, *Twin births and twins from a hereditary point of view*, Stockholm, A. B. Tidens Tryckeri.
9. Dahlberg Gunnar 1951, « Acta Genet. et Stat. Medica », II : 245.
10. Dahlberg Gunnar 1952, *Die Tendez zu Zwillingsgeburten* , « Acta Genet. Med. et Gemellologiae », I : 80.
11. Gedda, L. 1951, *Studio dei Gemelli*, Edizioni Orizzonte Medico, Roma, XVI : 1381.

12. GEISSLER, A. 1889, *Beitrage zur Frage des Geschlechts-verhältnisses der Geborenen*, « Zeits. K. Sachs. Statist. Bureaus », 35 : 1-24.

13. GINI, C. 1905, Doctor's Thesis, Univ. of Bologne.

14. GINI, C. 1908, *Il sesso dal punto di vista statistico*, Sandron, Palermo, (on sale at the library of 'Metron', Univ. of Rome).

15. GINI, C. 1911, *Considerazioni sulle probabilità a posteriori e applicazioni al rapporto dei sessi nelle nascite umane*, « Studi economico-giuridici della Università di Cagliari », Arin. III. (Republished « Metron », XV, in 1949).

16. GINI, C. 1951, *Combinations and Sequences of sexes in Human families and Mammal litters*, « Acta Genet. et Statlst. Med. », II : 220.

17. HAMILTON, W. J., J. D. BOYD and H. W. MOSSMAN 1947, *Human Embryology*, pp. 25, 65 & 73.

18. HERTWIG, R. 1912, « Biol. zentral blatt. », 32 : 1.

19. JENKINS, R. L. 1927, *The Interrelations of the frequency of plural births*, J. Hered. 8 : 387 & 504.

20. JENKINS, R. L. 1929, *Twin and Triplet birth ratios*, J. Hered. XX :485-494.

21. NEEDHAM, J. 1931, *Chemical Embriology*, Macmillan & Co. New York.

22. PRICE, Bronson 1950, *Primary Biases in Twin studies*, « Am. J. Human Genetics », 2 : 293-352.

23. SCHULTZ, A. H. 1918, *Studies in the Sex-ratio*, « Man Biol. Bull. », 34 : 257-275.

24. SLATER, E. 1944, *A Demographic study of the Psychopathic population*, « Ann. Eugen., Lond. », 2 : 2.

25. STOCKS, Percy, 1952, *Recent Statistics of Multiple births in England and Wales*, « Acta Genet. Med. et Gemellologiae », 1 : 8.

26. STRANDSKOV, H. H. 1945, *Multiple birth confinement frequencies in the U. S. birth registration area from* 1922-1936 *inclusive*, « Am. J. Phys. Anthrop. », 3 (n.s.) : 49.

27. ID. ID., and G. T. SIEMENS, 1946, *An Analysis of the sex-ratios among single and plural births in the total, the 'white' and the 'coloted' U.S. Populations*, « Am. J. Phys. Anthrop. », 4 (n.s.) : 491-501.

28. ID. ID., and H. BISACCIA, 1949, *The sex-ratio of human still births at each month of ffiterogestation and at conception*, « Am. J. Phys. Anthrop », 7 (n.s.) : 131.

29. TAUBER, R. 1923, Zeits. f. Guburtsch. U. Gynäk. vol. 85.

30. TIETZE, C. 1948, Human Blol. 20 : 156.

31. WEDERVANG, I. 1924, *Sex proportion and its variation in relation to antenatal mortality*, Oslo.

32. WINDLE, W. F. 1940, *Physiology of Foetus*, W. B. Saunders Co., Phila, and Lond.

33. PAPER I. Das, S. R. 1953, *A Mathematical Analysis of the Phenomena of Human Twins and Higher Plural births. Part I i Twins.* « Metron », vol. XVII - N. 1-2, Roma, Italia.

34. PAPER II. Das, S. R. 1953, *A Mathematical Analysis of the Phenomena of Human Twins and Higher Plural births. Part II : Triplets, and Application of the Analysis in the Intepretation of the Twin and the Triplet Data*, « Metron », vol. XVII - N. 3-4, Roma, Italia.